Sparse Linear Discriminant Analysis with more Variables than Observations

Thesis

Sparse Linear Discriminant Analysis with more Variables than Observations

by

Tsegay Gebrehiwot Gebru

(B.Sc. and M.Sc. in Statistics, Addis Ababa University)

A thesis submitted to The Open University

in fulfilment of the requirements for the degree of

Doctor of Philosophy in Statistics

School of Mathematics and Statistics

Faculty of Science, Technology, Engineering and Mathematics

The Open University

Walton Hall, Milton Keynes, MK7 6AA, United Kingdom

June 2018

Abstract

It is known that classical linear discriminant analysis (LDA) performs classification well when the number of observations is much larger than the number of variables. However, when the number of variables is larger than the number of observations, classical LDA cannot be performed because the within-group covariance matrix is singular. Recently proposed LDA methods that can handle singular within-group covariance matrix were reviewed. Most of these methods focus on regularizing the within-class covariance matrix. However, they give less attention to sparsity (selecting variables), interpretation and computational cost, which are important in high-dimensional problems. The fact that most of the original variables may be irrelevant or redundant suggests looking for sparse solutions that involve only a small portion of the variables. In the present work, new sparse LDA methods are proposed that are suited to high-dimensional data. The first two methods assume groups share a common within-group covariance matrix and approximate this matrix by a diagonal matrix. One of these methods is a variant of the other that sacrifices some accuracy for greater computational speed. Both methods obtain sparsity by minimizing an ℓ_1-norm and maximizing discrimination power under a common loss function with a tuning pa-

rameter. The third method assumes that groups share common eigenvector in eigenvector-eigenvalue decomposition of their within-group covariance matrices, while their eigenvalues my differ. The fourth method assumes the within-group covariance matrices are proportional to each other. The fifth method is derived from the Dantzig selector and uses optimal scoring to construct discriminant function. The third and fourth methods achieve sparsity by imposing a cardinality constraint with the cardinality level determined by cross-validation. All the new methods reduce their computation time by sequentially determining individual discriminant functions. The methods are applied to six real data sets and perform well when compared with two existing methods.

Acknowledgement

The accomplishment of this doctoral thesis would not have been possible without the support and encouragement of a number of people. I would like to express my sincere gratitude to all of them. First of all, I am extremely grateful to my PhD supervisors: Prof. Paul Garthwaite, Dr. Nickolay Trindafilov, and Prof. Frank Critchley who are staff members of the school of Mathematics and Statistics, The Open University. I thank Prof. Paul Garthwaite, my first supervisor, for his valuable guidance, scholarly inputs and consistent encouragement I received throughout the research work, particularly in the last one year. This accomplishment would not have been realized without his unconditional support and it was a great opportunity to do my doctoral programme under his guidance and to learn from his research expertise. I would also like to thank Dr. Nickolay Trindafilov, who was my first supervisor for the first 3 years of my PhD study, for all his guidance and unreserved scholarly supports in defining the research problem, in suggesting directions and positive inputs so as to make my study feasible theoretically and practically. I would also like to thank Prof. Frank Critchley for his guidance and support as my second supervisor. He gave me fruitful comments to shape my PhD research works and the thesis.

I extend my gratitude to all staff members of the Statistics group at the Open University. To mention some of them: Dr Alvaro Faria, Prof Chris Jones, Dr. Catriona Queen, Dr. Karen Vines, Dr. Heather Whitaker, Prof. Kevin McConway, Prof. Paddy Farrington, and Dr. Fadlalla Elfaday. They were very kind enough to extend their help at various phases of this research, whenever I approached them, and I do hereby acknowledge all of them. I would also like to thank the Open University for funding my PhD study.

This is good opportunity to thank Dr. Ian Short, senior lecturer of Mathematics at the Open University, for all his optimistic and continuous help and encouragements during the difficult time in my studies. Related with this, my thank also goes to Prof. Uwe Grimm, Head of the school of Mathematics and Statistics, for his help to realize the completion of my PhD.

I thank Dr. Yonas Weldeselassie for his friendly and brotherly support in various kinds from the beginning of the start of my PhD up to the completion of my PhD. I would also like to thank Saba Berhanu (wife of Dr Yonas) for her encouragements. Similarly, I would like to thank Dr. Yoseph Nugusse and his wife (Selam) for their advices and encouragements.

Last but not least, I would like to thank my family members, friends, classmates, officemates, and colleagues for their continuous encouragements throughout my PhD studies.

Contents

List of Tables

List of Figures

List of Abbreviations

CPC	Common Principal Components
DA	Discriminant Analysis
FC-SLDA	Function Constrained Sparse Linear Discriminant Analysis
LDA	Linear Discriminant Analysis
LDF	Linear Discriminant Function
ODE	Ordinary Differential Equations
OS	Optimal Scoring
PCA	Principal Components Analysis
PLDA	Penalized Linear Discriminant Analysis
QDA	Quadratic Discriminant Analysis
SDA	Sparse Discriminant Analysis
SDCPC	Sparse Discrimination with CPC
SD-PCPC	Sparse Discrimination with Proportional CPC

SLDA-OS Sparse Linear Discriminant Analysis with Optimal Scoring

SVD Singular Value Decomposition

Chapter 1

Introduction and preliminaries

1.1 Introduction

With the recent development of new technologies, high-dimensionality has become a common problem in various disciplines such as medicine and epidemiology, genetics, biology, metrology, astronomy, and economics. High-dimensionality is a situation where the number of variables (the dimension of the data vectors) is much larger than the number of observations (sample size) (Qiao et al., 2009). Some sources of high-dimensional data are digital images, documents, next-gen sequencing, mass spectrometry, metabolomics, microarray (gene expression), proteomics, videos and web pages (Pang and Tong, 2012). The high-dimensionality problem, in general, occurs in many applications including information retrieval, character recognition, classification and microarray data analysis (Ye, 2005).

To analyse high-dimensional data, many methods have been proposed for fast query response, such as K-D tree and R-tree (Cai et al., 2008). However, the

performance and efficiency of these types of methods decrease as the dimensionality increases because the methods are designed to operate with small dimensionality. Consequently, dimension reduction, or variable selection, has become an important approach to deal with high-dimensional problems so as to obtain meaningful results. Once the high-dimensional data are transformed into a lower dimensional space, conventional data analysis methods can be employed (Cai et al., 2008). One of the most commonly used dimension reduction methods for data with grouped observations is Discriminant Analysis (DA). Principal Component Analysis (PCA) is another popular method used for dimension reduction. It helps to find a few directions on which to project the data such that the projected data explain most of the variability in the original data. This method finds a low dimensional representation of the data without losing much information. Although PCA can be used for dimension reduction, it is not appropriate for classification problems because it mainly works for unsupervised problems (Qiao et al., 2009).

Discriminant Analysis (DA) is generally defined as the study of the relationship between a categorical variable and a set of interrelated variables (McLachlan, 2004). A method that is commonly used together with DA is classification, which is a supervised method that deals with the problem of the optimal allocation of a given set of objects into a predefined mutually exclusive and exhaustive classes. Fisher (1936) proposed a special type of DA called linear discriminant analysis (LDA). It is a method used in statistics, pattern recognition and machine learning to find a linear combination of variables, linear discriminant functions

(LDF), which characterize or separate two or more groups of objects or events. The resulting linear combination of variables may be used as a classifier, or, more commonly, for dimensionality reduction before classification. The main objective of LDA is to describe, either graphically (in few directions) or algebraically, the difference between two or more groups of objects as well as to perform dimensionality reduction while preserving as much of the class discriminatory information as possible (Johnson and Wichern, 2002).

The classical Fisher's LDA approach uses the class information to find informative projections of the data for a classification problem. Fisher (1936) considered the problem of finding a linear combination of variables that best discriminates groups by maximizing the ratio of between-class variance to within-class variance. In the case of two classes, the derived linear combination of variables is called a linear discriminant function (LDF), or canonical variate (Trendafilov and Vines, 2009). In the same manner, additional LDF's with decreasing importance in discrimination can be obtained sequentially (Qiao et al., 2009). This method of discrimination is further generalized by Rao (1952) to the multiple class problem. In general, when the number of variables is greater than the number of groups, the total number of discriminant functions that can be defined is one less than the number of groups. For example, when there are three groups, we could estimate two discriminant functions, one function for discriminating between group 1 and groups 2 and 3 combined, and another function for discriminating between group 2 and group 3.

Another way of deriving LDA originates from the assumption that each class

follows a multivariate normal distribution with significantly different group means but a common covariance matrix (Merchante et al., 2012; Trendafilov and Jolliffe, 2007; Johnson and Wichern, 2002). Together with the minimization of the probabilities of misclassification, this basic normality assumption leads to a Bayes discrimination method that coincides with Fisher's LDA. Alternatively, Fisher's LDA can also be formulated as a linear regression model through the concept of optimal scoring of the classes (Mai et al., 2012; Clemmensen et al., 2011; Merchante et al., 2012; Hastie et al., 1995).

It is well known that classical LDA is one of the dimension reduction methods that performs well when the number of observations to be classified is much larger than the number of variables used for discrimination and classification. However, in the high dimensional setting, that is, when the number of variables is much larger than the number of observations, classical LDA fails to perform classification effectively due to the following well known problems (Clemmensen et al., 2011; Fan et al., 2012; Witten and Tibshirani, 2011; Ng et al., 2011; Hastie et al., 1995).

1. The estimate of the within-group covariance matrix is singular.

2. The resulting discriminant functions are very difficult to interpret, because each discriminant function includes a linear combination of all of the original variables.

3. Computational cost in terms of both running time and storage is very expensive.

Furthermore, many more problems of high-dimensional data have been identified by various studies. For instance, Bickel and Levina (2004) pointed out that Fisher's LDA performs poorly in a minimax sense due to the *diverging spectra* frequently encountered in high-dimensional covariance matrices. Fan and Fan (2008) also demonstrated that the difficulty in high-dimensional classification is due to the presence of redundant variables (*noise accumulation*) that do not significantly contribute to the minimization of classification error or to the maximization of discrimination between groups. Similarly, Qiao et al. (2009) stated that in high-dimensional discriminant analysis, most of the time data are projected onto various directions, many of the projections are exactly the same. That is, the data overlap on top of each other. They referred to this phenomenon as *data pilling* or over fitting.

In general, many effective statistical techniques such as LDA cannot even be computed directly in high-dimensional data due to the aforementioned problems. If LDA is directly applied to such data settings, it may provide meaningless results. Therefore, appropriate methods of transformation or dimension reduction are required to apply LDA in such circumstances.

There exist several references that have proposed various methods to extend classical LDA to overcome the problems that arise in the high-dimensional setting. Recently proposed extensions of LDA focus mainly on dimension reduction through variable selection and on the estimation of the inverse of the within-class covariance matrix by applying different regularization techniques (Clemmensen et al., 2011; Witten and Tibshirani, 2011; Qiao et al., 2009; Fan et al., 2012; Fan and

Fan, 2008; Ng et al., 2011).

Variable selection is an approach by which high-dimensional data is reexpressed in terms of fewer variables while minimizing the loss of necessary information for discrimination (Merchante et al., 2012; Hastie et al., 1995). The variables obtained after the final dimension reduction process are commonly called discriminant variables (Hastie et al., 1995). The main purpose of variable selection is to achieve sparsity. Sparsity is a situation where the discriminant vectors have only a small number of nonzero components (Qiao et al., 2009). In other words, sparse LDA produces linear discriminant functions with only a small number of variables, retaining those variables that are important in discriminating between groups and in identifying group membership of observations. In high-dimensional data analysis, such as most genetic analyses, sparse methods of discrimination ensure better interpretability, robustness of the model, or less computational cost for prediction (Clemmensen et al., 2011; Merchante et al., 2012).

Variable selection is an essential procedure in the derivation of sparse LDA. In high-dimensional data, often a large number of variables on which measurements are observed are available for analysis, while few of these variables contain useful information for the purpose of classification (Rencher, 2002). Qiao et al. (2009) pointed out that we do not necessarily ensure an increase in the discriminatory power by increasing the number of variables in the application of Fisher's LDA. Instead it leads to formation of overfitting. Since the 1990's, a number of techniques have been proposed for variable selection with high-dimensional

data. The prominent methods are variable selection via the Lasso (Tibshirani, 1996), variable selection via the elastic net (Zou and Hastie, 2005), the Dantzig selector (Candès and Tao, 2007), and the group Lasso (Merchante et al., 2012). The traditional approach to sparse LDA is performing variable selection in a separated step before classification. However, this approach leads to a dramatic loss of information for the purpose of the overall classification problem (Filzmoser et al., 2012). Therefore, there is a need to develop a sparse LDA method that performs variable selection and classification simultaneously.

1.2 Thesis outline

Each of the chapters in this thesis can be read as a self-contained article. In general, the thesis is organized as follows. Chapter 2 briefly introduces the general discriminant analysis framework. Various techniques of classical discriminant analysis are presented to give a general background about discriminant analysis. The principles of classification and discrimination are presented here. Moreover, three approaches to discriminant analysis are presented in this chapter. These are discrimination via multivariate normal models, Fisher's LDA, and the regression approach to LDA.

Chapter 3 reviews some of the existing discriminant approaches in high dimensional settings. This chapter, in general, reviews the approaches that focus on dimension reduction, regularization of the within-groups sample covariance matrix, minimization of classification error, and other direct methods. With these approaches, ordinary LDA is used after dimension reduction. Other methods

that are reviewed in Chapter 3 are methods that assume the variables in a high-dimensional data are independent.

Chapter 4 proposes a method called function-constrained sparse LDA (FC-SLDA) and its simplified version, FC-SLDA2, that are alternative methods for high-dimensional discriminant analysis. The constrained ℓ_1-minimization penalty is imposed on the discrimination problem to achieve sparsity, and FC-SLDA imposes a diagonal within-group covariance matrix to circumvent the singularity problem. The second method proposed in this chapter, FC-SLDA2, is derived without using eigenvalues. Both methods are illustrated using real data sets. They are also compared with other exiting methods.

Chapter 5 starts by introducing a new method of discrimination called sparse LDA using Common principal components (CPC) and then continues with the theoretical development of the method. Sparse discriminant method using CPC (SDCPC) assumes that group covariance matrices have the same eigenvectors but different eigenvalues. It is an effective method for high-dimensional classification problems. This method is illustrated by using real data sets. Finally, Chapter 5 proposes another alternative sparse discrimination method called sparse LDA using proportional CPC (SD-PCPC) for high-dimensional discrimination problems. This method is appropriate when group covariance matrices are proportional to each other.

Chapter 6 proposes a new formulation to sparse LDA method based on optimal scoring named as SLDA-OS. This discrimination method is derived by recasting discriminant analysis as regression analysis. The Danzig selector is incor-

porated within this method to achieve sparsity of the discriminant functions.

The thesis ends with summary and conclusions in Chapter 7, where each chapter is briefly summarized, results are discussed, and conclusions are presented. Some future research directions are also indicated in this chapter. We used MATLAB2015b to implement the algorithms of our methods.

Chapter 2

The discriminant analysis framework

In this chapter, we outline the general framework (formulation) of the discrimination problem and present the main approaches of classical discriminant analysis.

2.1 Discrimination and classification problems

Discriminant analysis and classification are multivariate techniques concerned with separating distinct sets of objects and with allocating new objects to previously defined groups. Discriminant analysis is a dimension reduction method that is useful in determining whether a set of variables is effective in predicting group membership. For example, linear discriminant analysis (LDA) is used to identify a linear combination of variables, called the linear discriminant function, that produces the greatest distance between groups. A restriction on using standard LDA is that it requires group covariances to be equal. Some other nonlinear discriminant analysis, such as quadratic discriminant analysis (QDA), may be used when the group covariances are not equal.

The goal of discrimination, in general, is to describe the differential features of objects that can be used to separate the objects into groups as well as to predict group membership of further objects (Fisher, 1936). The latter task overlaps classification analysis which is concerned with the development of rules for allocating or assigning observations into one or more already existing groups.

Because linear discriminant functions are often used to develop classification rules, some authors use the term classification analysis instead of discriminant analysis. Because of the close association between the two processes we treat them together in this subsection.

2.2 Basic notation and data organization

Multivariate data for discriminant analysis arise when measurements made on p variables are recorded for a total of n observations (individuals). Because we are now dealing with classical LDA, we assume that $n > p$. Suppose that the n observations are divided into g predefined groups and that the i^{th} group is denoted by π_i, $i = 1, 2, \ldots, g$. If n_i is the number of observations in the i^{th} group, then $n_1 + n_2 + \cdots + n_g = n$. Let the $(p \times 1)$ vector $\mathbf{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})^T$ denote the measurement made on the j^{th} individual belonging to the i^{th} group, and let the $(n \times p)$ data matrix $\mathbf{X}$ represent the measurements of all observations. Values will be available for p variables $X_1, X_2, ..., X_p$ for each observation. Thus, the data for discriminant analysis takes the form shown in Table 2.1.

Therefore, the matrix $\mathbf{X}$ contains the data consisting of all of the n observations on all of the p variables in g groups. It can also be given as $\mathbf{X}^T =$

Table 2.1: *Multivariate data for discriminant analysis*

Observation	X_1	X_2	...	X_p	Group
1	x_{111}	x_{112}	...	x_{11p}	1
2	x_{211}	x_{212}	...	x_{21p}	1
⋮	⋮	⋮	⋮	⋮	⋮
n_1	$x_{n_1 11}$	$x_{n_1 12}$	...	$x_{n_1 1p}$	1
1	x_{121}	x_{122}	...	x_{12p}	2
2	x_{221}	x_{222}	...	x_{22p}	2
⋮	⋮	⋮	⋮	⋮	⋮
n_2	$x_{n_2 21}$	$x_{n_2 22}$	...	$x_{n_2 2p}$	2
⋮	⋮	⋮	⋮	⋮	⋮
1	x_{1g1}	x_{1g2}	...	x_{1gp}	g
2	x_{2g1}	x_{2g2}	...	x_{2gp}	g
⋮	⋮	⋮	⋮	⋮	⋮
n_g	$x_{n_g g1}$	$x_{n_g g2}$	...	$x_{n_g gp}$	g

$[X_1, X_2, ..., X_p]$.

2.3 Principles of classification and discrimination

2.3.1 Classification into two groups

Suppose the overall set of measurements on n observations is divided into two groups. The first group is π_1 and contains n_1 observations; the second group π_2 contains n_2 observations. Let these two populations be described by probabil-

ity density functions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$, respectively, where the observed values of $\mathbf{x}$ differ to some extent from one group to the other (Johnson and Wichern, 2002).

An observation with associated measurements $\mathbf{x}$, must be assigned to either π_1 or π_2. Let Ω be the sample space; that is, the collection of all possible observations $\mathbf{x}$. The space is divided into two regions, say, R_1 and $R_2 = \Omega - R_1$. If an observation falls in R_1, we classify it as belonging to π_1, and if the observation falls in R_2, we classify it as belonging to π_2. Since every observation must be assigned to one and only one of the two populations, the regions R_1 and R_2 are mutually exclusive and exhaustive (Johnson and Wichern, 2002).

In using any classification procedure, two types of errors can be committed: an observation may be incorrectly classified as coming from π_2 when, in fact, it is from π_1, and viceversa (Anderson, 1984). The principle of optimal allocation is to create a rule (R_1 and R_2) that minimizes the chances of making these errors. In general, a large number of observations tend to be classified into their respective groups.

With good classification method, the chances or probabilities of misclassification should be small. The conditional probability of classifying an object as π_2 when , in fact, it is from π_1 is given as :

$$p(2|1) = p(\mathbf{X} \in R_2|\pi_1) = \int_{R_2} f_1(\mathbf{x})d\mathbf{x} \tag{2.1}$$

Similarly, the conditional probability of classifying an object as π_1 when it is really from π_2 is

$$p(1|2) = p(\mathbf{X} \in R_1|\pi_2) = \int_{R_1} f_2(\mathbf{x})d\mathbf{x} \tag{2.2}$$

Let p_i be a prior probability of π_i $(i = 1, 2)$, where $p_1 + p_2 = 1$. Therefore, the

overall probabilities of correctly or incorrectly classifying objects can be derived as the product of the prior and conditional classification probabilities.i.e.,

$$p(\text{correctly classified as } \pi_1) = p(\mathbf{X} \in R_1|\pi_1).p(\pi_1) = p(1|1).p_1 \tag{2.3}$$

and

$$p(\text{misclassified as } \pi_1) = p(\mathbf{X} \in R_1|\pi_2).p(\pi_2) = p(1|2).p_2 \tag{2.4}$$

In the same manner, the probabilities of correctly and incorrectly classifying observations as π_2 are given as , respectively:

$$p(\mathbf{X} \in R_2|\pi_2).p(\pi_2) = p(2|2).p_2 \tag{2.5}$$

and

$$p(\mathbf{X} \in R_2|\pi_1).p(\pi_1) = p(2|1).p_1. \tag{2.6}$$

Classification methods are often evaluated based on their probabilities of misclassification (PoM). A classification procedure with smaller PoM is said to be better than another method of classification with larger PoM. Consequently, in the case of two groups classification process, the idea of classification is to develop a method that minimizes the PoM's in equations 2.4 and 2.6.

Another criteria for classification is cost. Suppose that classifying a π_1 observation wrongly to π_2 represents a more severe error than classifying a π_2 observation wrongly to π_1. Then one should be cautious about committing the former error. Let the cost of an observation from π_1 is misclassified as π_2 be $c(2|1)$, and the cost of an observation from π_2 is misclassified as π_1 be $c(1|2)$. Then the average or expected cost of misclassification (ACM) is given as:

$$\text{ACM} = c(2|1).p(2|1).p_1 + c(1|2).p(1|2).p_2. \tag{2.7}$$

It is noted in Johnson and Wichern (2002) that the cost for correct classification is zero. A reasonable classification rule aims to have an ACM as small as possible.

2.3.2 Optimal allocation criteria

Many different optimal allocation criteria have been proposed to determine a classification rule. One criterion is to obtain a classification rule by minimizing the ACM. A procedure that minimizes (2.7) for given p_1 and p_2 is called a Bayes rule (Anderson, 1984). The regions R_1 and R_2 that minimize the ACM are defined by the values x for which the following inequalities hold

$$R_1 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right), \tag{2.8}$$

and

$$R_2 : \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right). \tag{2.9}$$

If the misclassification cost ratio is unknown, it is commonly taken to be unity and the population density ratio is compared with the ratio of the prior probabilities. Suppose for a moment that $c(1|2) = c(2|1) = 1$. Then the expected cost of misclassification (the ACM) given in (2.7) becomes solely a function of the probabilities. As a result, we call it the total probability of misclassification (TPM), given as :

$$\begin{aligned}
\text{TPM} &= p_1.p(2|1) + p_2.p(1|2) \\
&= p_1 \int_{R_2} f_1(\mathbf{x})d\mathbf{x} + p_2 \int_{R_1} f_2(\mathbf{x})d\mathbf{x} \\
&= p_1 \left[1 - \int_{R_1} f_1(\mathbf{x})d\mathbf{x}\right] + p_2 \int_{R_1} f_2(\mathbf{x})d\mathbf{x} \\
&= p_1 + \int_{R_1} \left[p_2 f_2(\mathbf{x}) - p_1 f_1(\mathbf{x})\right] d\mathbf{x}.
\end{aligned} \tag{2.10}$$

This quantity is minimized if R_1 is chosen so that $p_2 f_2(\mathbf{x}) - p_1 f_1(\mathbf{x}) < 0$ for all points in R_1. Minimizing (2.10) is mathematically equivalent to minimizing the expected cost of misclassification when the costs of misclassification are equal (Johnson and Wichern, 2002). The classification rule that minimizes TPM is given as follows:

Assign an observation $\mathbf{x}$ to π_1 if

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1}; \tag{2.11}$$

otherwise assign it to π_2. Moreover, when the prior probabilities are unknown, they are taken to be equal, i.e., $p_1 = p_2 = 1/2$. Under both conditions, the optimal classification regions are determined simply by comparing the values of the density functions. Hence, with the assumption of equal cost of misclassification and equal prior probabilities, we assign an observation $\mathbf{x}$ to π_1 if $f_1(\mathbf{x})/f_2(\mathbf{x}) \geq 1$, otherwise we assign it to π_2.

Another optimality criterion that leads to the assignment rule in (2.11) is based on posterior probability. Using this approach an observation $\mathbf{x}$ is allocated to the group with the largest posterior probability $p(\pi_i|\mathbf{x})$. By Bayes rule, the posterior probability of π_i is given as:

$$\begin{aligned} p(\pi_i|\mathbf{x}) &= \frac{p(\mathbf{x}|\pi_i)p(\pi_i)}{\sum_{k=1}^{2} p(\mathbf{x}|\pi_k)p(\pi_k)} \\ &= \frac{p_i f_i(\mathbf{x})}{p_1 f_1(\mathbf{x}) + p_2 f_2(\mathbf{x})}, \ i = 1, 2. \end{aligned} \tag{2.12}$$

An observation $\mathbf{x}$ is assigned to π_1 when $p(\pi_1|\mathbf{x}) > p(\pi_2|\mathbf{x})$, this is equivalent to the rule that minimizes the total probability of misclassification.

An alternative criterion specifies that the maximum probability of misclassi-

fication should be minimized. This criterion is commonly known as the *minimax* rule. Thus, the minimax rule allocates an observation $\mathbf{x}$ so as to minimize the greater of $p(1|2)$ and $p(2|1)$ (Lachenbruch, 1975; Seber, 2004). For instance, for $0 \leq \alpha \leq 1$,

$$\max\{p(1|2), p(2|1)\} \geq (1-\alpha)p(2|1) + \alpha p(1|2) \tag{2.13}$$

By (2.11) the right hand side of (2.13) is minimized when $R_1 = R_{01} = \{f_1(\mathbf{x})/f_2(\mathbf{x}) \geq \alpha/(1-\alpha) = c\}$. If we choose c, say $\alpha = \alpha_0$, so that the misclassification probabilities for R_{01} are equal, that is, $p_0(2|1) = p_0(1|2)$, then

$$\begin{aligned}(1-\alpha_0)p(2|1) + \alpha_0 p(1|2) &\geq (1-\alpha_0)p_0(2|1) + \alpha_0 p_0(1|2) \\ &= (1-\alpha_0+\alpha_0)p_0(2|1) \\ &= p_0(2|1)\end{aligned}$$

Therefore, (2.13) can be given as,

$$\max\{p(1|2), p(2|1)\} \geq p_0(2|1) = \max\{p_0(1|2), p_0(2|1)\}.$$

Thus, the minimax rule is: Assign $\mathbf{x}$ to π_1 if $f_1(\mathbf{x})/f_2(\mathbf{x}) \geq$ c, where c satisfies $p_0(1|2) = p_0(2|1)$.

If the two groups are normal with common covariance matrix, then the minimax rule is given as: Assign an observation $\mathbf{x}$ to π_1 if

$$D(\mathbf{x}) \geq \ln c,$$

where $D(\mathbf{x})$ is given by (2.21). The minimax rule is the same as the maximum likelihood ratio method when $\ln c = 0$ or $c = 1$. Both allocation methods do not require knowledge of p_1.

2.3.3 Classification into several groups

Here the principles of classification presented in the previous sections will be extended to the case where there are more than two groups. Let the observations be divided into g groups, where the i^{th} group is denoted by π_i with associated density functions $f_i(\mathbf{x})$, $i = 1, 2, \ldots, g$. The space of observations is assumed to be divided into g mutually exclusive and exhaustive regions $R_1, R_2, \ldots, R_g$. Let p_i be the prior probability of π_i, and let $c(k|i)$ be the cost of assigning an observation wrongly to π_k when, in fact, it belongs to π_i for $i \neq k = 1, 2, \ldots, g$. For $k = i, c(i|i) = 0$. Similarly, let $p(k|i)$ be the probability of misclassifying an observation to π_k when, in fact, it comes from π_i, which is given as:

$$p(k|i) = \int_{R_k} f_i(\mathbf{x}) d\mathbf{x} \quad \text{for} \quad i, k = 1, 2, \ldots, g. \tag{2.14}$$

with

$$p(i|i) = 1 - \sum_{\substack{k=1 \\ k \neq i}}^{g} p(k|i)$$

The conditional expected cost of misclassifying an observation $\mathbf{x}$ from π_1 to π_2 or $\pi_3, \ldots,$ or π_g is:

$$\begin{aligned} ACM(1) &= p(2|1)c(2|1) + p(3|1)c(3|1) + \cdots + p(g|1)c(g|1) \\ &= \sum_{k=2}^{g} p(k|1)c(k|1). \end{aligned} \tag{2.15}$$

The conditional expected costs of misclassification for the other groups can also be obtained from equivalent formula. Multiplying each conditional expectation

by its prior probability and summing the results gives the overall ACM:

$$ACM = \sum_{i=1}^{g} p_i \left(\sum_{k=1, k\neq i}^{g} p(k|i)c(k|i) \right). \tag{2.16}$$

Determining an optimal classification procedure means choosing $R_K, k = 1, 2, \ldots, g$ so that (2.16) is minimized. The allocation rule is: Assign $\mathbf{x}$ to $\pi_k, k = 1, 2, \ldots, g$ for which (2.16) is smallest (Johnson and Wichern, 2002). If all the misclassification costs are equal, the minimum ACM and the minimum TPM are the same and, without loss of generality, we can set all the misclassification costs equal to 1. This assumption leads to the allocation rule that we would allocate $\mathbf{x}$ to group $\pi_k, k = 1, 2, \ldots, g$, for which

$$\sum_{i=1, i\neq k}^{g} p_i f_i(\mathbf{x}) \tag{2.17}$$

is smallest. Note that equation (2.17) will be smallest when the omitted term, $p_k f_k(\mathbf{x})$, is largest. As a result, when all the misclassification costs are the same, the allocation rule is that we assign $\mathbf{x}$ to π_k if $p_k f_k(\mathbf{x}) > p_i f_i(\mathbf{x})$ for all $i \neq k$. It is important to note that this classification rule is identical to the one that maximizes the posterior probability $p(\pi_k|\mathbf{x})$, where

$$p(\pi_k|\mathbf{x}) = \frac{p_k f_k(\mathbf{x})}{\sum_{i=1}^{g} p_i f_i(\mathbf{x})} \quad \text{for } k = 1, 2, \ldots, g. \tag{2.18}$$

Equation (2.18) is the generalization of equation (2.12) for g groups.

2.4 Approaches to linear discriminant analysis

There are many approaches to LDA. In this section, we will present three approaches, namely, multivariate normal discrimination, Fisher's discrimination, and discrimination using regression approach.

2.4.1 Discrimination via multivariate normal models

2.4.1.1 Discrimination with two multivariate normal populations

Here we assume that $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are multivariate normal densities; the first with mean vector $\boldsymbol{\mu}_1$ and covariance matrix $\boldsymbol{\Sigma}$, and the second with mean vector $\boldsymbol{\mu}_2$ and the same covariance matrix, $\boldsymbol{\Sigma}$. We also assume that all of the population parameters are known. The multivariate normal density of $\mathbf{x}^T = [x_1, x_2, ..., x_p]$ for the i^{th} group is:

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)], \quad i = 1, 2. \tag{2.19}$$

Thus, the ratio of the densities is:

$$\begin{aligned} \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= \frac{\exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)]}{\exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)]} \\ &= \exp[(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)] \end{aligned} \tag{2.20}$$

Taking logarithm the optimal rule becomes: Assign $\mathbf{x}$ to π_1 if

$$D(\mathbf{x}) = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)) > \ln\frac{p_2}{p_1}; \tag{2.21}$$

otherwise assign $\mathbf{x}$ to π_2. Note that the inequality in (2.21) is found when the costs of misclassification are assumed to be equal. Moreover, when $p_1 = p_2 = 1/2$, $\mathbf{x}$ will be assigned to π_1 if $D(\mathbf{x}) > 0$.

$D(\mathbf{x})$ can be rewritten as:

$$D(\mathbf{x}) = \mathbf{w}^T(\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)) \tag{2.22}$$

where $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$. It is important to see that $D(\mathbf{x})$ is a linear function of the observation vector $\mathbf{x}$ and hence it is known as the linear discriminant func-

tion (LDF). In fact, $\mathbf{w}^T$ in (2.22) is a row vector which can be given as, $\mathbf{w}^T = (w_1, w_2, \ldots, w_p)$. For example, if an observation $\mathbf{x_0}$ consists of $(x_{01}, x_{02}, \ldots, x_{0p})$, then the discriminant score, $D(\mathbf{x})$, is computed as:

$$D(\mathbf{x_0}) = w_0 + w_1 x_{01} + w_2 x_{02} + \cdots + w_p x_{0p}$$

where w_0 is a constant given by $w_0 = \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$ (Rencher, 2002; Johnson and Wichern, 2002).

When $p_1 = p_2 = 1/2$, we assign $\mathbf{x}$ to π_1 if

$$\mathbf{w}^T\mathbf{x} \geq \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = \frac{1}{2}(\mathbf{w}^T\boldsymbol{\mu}_1 + \mathbf{w}^T\boldsymbol{\mu}_2) \tag{2.23}$$

This means that we assign $\mathbf{x}$ to π_1 if $\mathbf{w}^T\mathbf{x}$ is closer to $\mathbf{w}^T\boldsymbol{\mu}_1$ than to $\mathbf{w}^T\boldsymbol{\mu}_2$.

To find the probabilities of misclassification, it is useful to know the distribution of $D(\mathbf{x})$. First, let us define the squared Mahalanobis distance between $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ as

$$\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathbf{w}^T(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{2.24}$$

The distribution of $D(\mathbf{x})$ is derived as follows. Since $\mathbf{x}$ is multivariate normal, $D(\mathbf{x})$ is also normal. This is because $D(\mathbf{x})$ is a linear combination of $\mathbf{x}$. If $\mathbf{x}$ comes from $\pi_i (i = 1, 2)$, the mean of $D(\mathbf{x})$ is

$$\begin{aligned} E[D(\mathbf{x})|\pi_i] &= E[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x_i} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2))] \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)) \\ &= \frac{1}{2}(-1)^{i+1}\Delta^2. \end{aligned} \tag{2.25}$$

In either population the variance (var) is

$$\begin{aligned} \text{var}(D(\mathbf{x})) &= \text{var}(\mathbf{w}^T(\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2))) \\ &= \text{var}(\mathbf{w}^T\mathbf{x}) = \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w} = \Delta^2. \end{aligned} \tag{2.26}$$

Thus the probability of misclassification if the observation is from π_1 is $p(2|1) = \Phi(\{\ln \frac{p_2}{p_1} - \frac{\Delta^2}{2}\}/\Delta)$, where $\Phi(.)$ denotes the standard normal distribution function. Similarly, $p(1|2) = \Phi(-\{\ln \frac{p_1}{p_2} + \frac{\Delta^2}{2}\}/\Delta)$. If we assume that $p_1 = p_2 = 1/2$, then

$$p(2|1) = p(1|2) = \Phi\left(-\frac{\Delta}{2}\right); \tag{2.27}$$

and the total probability of misclassification is given as

$$\begin{aligned} TPM &= \frac{1}{2}p(D(\mathbf{x}) < 0|\pi_1) + \frac{1}{2}p(D(\mathbf{x}) > 0|\pi_2) \\ &= \frac{1}{2}\Phi\left(-\frac{\Delta}{2}\right) + \frac{1}{2}\Phi\left(-\frac{\Delta}{2}\right) \\ &= \Phi\left(-\frac{\Delta}{2}\right) = 1 - \Phi\left(\frac{\Delta}{2}\right). \end{aligned} \tag{2.28}$$

The allocation principle is to find a classifier $D(\mathbf{x})$ that minimizes the total probability of misclassification in (2.28).

If the assumption of equal population variances was violated, the function would be a quadratic discriminant function (details are given in Anderson (1984)). In such circumstances, quadratic discriminant analysis controls the variability in each group and provides reliable results.

2.4.1.2 Discrimination with several multivariate normal populations

Let $f_i(\mathbf{x})$ be a multivariate normal density function of $\mathbf{x}$ for population π_i with mean vector $\mu_\mathbf{i}$ and covariance matrix $\mathbf{\Sigma_i}, i = 1, 2, \cdots, g$. Let $\mathbf{\Sigma}$ be the common covariance matrix of the g populations under the assumption of homoscedasticity. The multivariate normal density of $\mathbf{x}$ for the i^{th} population is

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)\right], \quad i = 1, 2, \cdots, g \tag{2.29}$$

Multiplying by p_i and taking logarithm gives

$$D_i(\mathbf{x}) = \ln p_i f_i(\mathbf{x}) = \ln p_i - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) \tag{2.30}$$

Thus we assign $\mathbf{x}$ to π_k if $D_k(\mathbf{x}) = \max \ln p_i f_i(\mathbf{x}), i = 1, 2, \ldots, g$. The constant term $(\frac{p}{2}) \ln(2\pi)$ in (2.30) is the same for all groups. Hence, it can be ignored for allocation purposes. Similarly, we can ignore other terms that are the same for each $D_i(\mathbf{x})$. Consequently, the final linear discriminant score for the i^{th} group can be defined as

$$\begin{aligned} D_i(\mathbf{x}) &= \ln p_i + \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i \\ &= \ln p_i + \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_i) \end{aligned} \tag{2.31}$$

We assign $\mathbf{x}$ to the group with the largest value of $D_i(\mathbf{x})$.

It is important to note that the linear discriminant scores in g groups can also be expressed as:

$$D_{ik}(\mathbf{x}) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_k)) \quad \text{for } i, k = 1, 2, \ldots, g, \text{ and } i \neq k. \tag{2.32}$$

where $D_{ik}(\mathbf{x})$ is the discriminant function related to the i^{th} and k^{th} groups, and $D_{ik}(\mathbf{x}) = -D_{ki}(\mathbf{x})$. The region R_i is bounded by a $(g-1)$ dimensional hyperplane. The mean and variance of $D_{ik}(\mathbf{x})$ are, respectively, $\frac{1}{2}\Delta_{ik}^2$ and Δ_{ik}^2, where Δ_{ik}^2 is given as

$$\Delta_{ik}^2 = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_k) = \mathbf{w}_{ik}^T (\boldsymbol{\mu}_i - \boldsymbol{\mu}_k), \tag{2.33}$$

where $\mathbf{w}_{ik} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_k)$.

When the population parameters are unknown, they can be replaced by their sample counterpart plug-in estimates. The sample mean vector and covariance

matrix for the i^{th} $(i = 1, 2, ..., g)$ group are given by

$$\begin{aligned} \hat{\boldsymbol{\mu}}_i &= \bar{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{x}_{ij} \\ \hat{\boldsymbol{\Sigma}}_i &= \mathbf{S}_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T \end{aligned} \tag{2.34}$$

Similarly, $\boldsymbol{\Sigma}$ may be estimated by the pooled sample covariance which is given by

$$\hat{\boldsymbol{\Sigma}} = \mathbf{S} = \frac{\sum_{i=1}^{g}(n_i - 1)\mathbf{S}_i}{n - g} \tag{2.35}$$

And the overall mean vector $\boldsymbol{\mu}$ is estimated as:

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{g}\sum_{j=1}^{n_i}\mathbf{x}_{ij} \tag{2.36}$$

2.4.2 Fisher's linear discriminant analysis

Fisher's approach does not assume normality, but it assumes that the populations have equal covariance matrices. A pooled estimate of the covariance matrix will be used in this section. Similarly, sample estimates of the mean vectors will be used here.

It is convenient to start with two groups. Fisher (1936) determined the linear combination of p variables

$$y = \mathbf{a}^T\mathbf{x} = a_1x_1 + a_2x_2 + ... + a_px_p \tag{2.37}$$

that maximizes the distance between the two group mean vectors. This linear combination transforms the multivariate observations $\mathbf{x}$ to univariate observations (scalar) y such that the $y's$ derived from populations π_1 and π_2 are separated as much as possible. The objective is to find the vector $\mathbf{a}$ that maximizes the

standardized distance between the two group means, which is given as

$$\frac{[\mathbf{a}^T(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)]^2}{\mathbf{a}^T\mathbf{S}\mathbf{a}} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) = \hat{\Delta}^2. \tag{2.38}$$

The maximum of (2.38) is obtained when $\mathbf{a} = \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, or when it is any multiple of $\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ (Rencher, 2002). Consequently, the linear combination in (2.37) can be rewritten as $y = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{S}^{-1}\mathbf{x}$. This function is called Fisher's linear discriminant function. It is identical to the standard LDA function ($\mathbf{w}^T\mathbf{x}$) with $\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))$ given in Section 2.4.1.1, but with unknown population means replaced by sample means. It can be shown that the maximizing vector $\mathbf{a}$ is not unique because any multiple of $\mathbf{a} = \mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ will maximize (2.38). However, its direction is unique. (Rencher, 2002, p.271).

To extend Fisher's approach of LDA to g groups, we first need to define the between-group and within-group covariance matrices. Let $\mathbf{B}$ be the between-groups covariance matrix and $\mathbf{W}$ be the within-groups covariance matrix of $\mathbf{X}$, which are given by

$$\mathbf{B} = \hat{\boldsymbol{\Sigma}}_b = \frac{1}{g-1}\sum_{i=1}^{g} n_i(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T$$

and

$$\mathbf{W} = \hat{\boldsymbol{\Sigma}}_w = \frac{1}{n-g}\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)^T. \tag{2.39}$$

We seek a linear combination of the original variables that transforms the p-dimensional vector $\mathbf{x}$ to s-dimensional vector $\mathbf{y}$, with $s < p$. The linear combination may be given as $\mathbf{y} = \mathbf{A}^T\mathbf{x}$, where $\mathbf{A}$ is a $p \times s$ transformation matrix that gives the greatest discrimination between groups by maximizing the ratio of the between-groups covariance matrix to the within-groups covariance matrix of

the data (Trendafilov and Vines, 2009). Suppose $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_s$ be respectively the s column vectors of the transformation matrix $\mathbf{A}$. The vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_s$ are obtained from $\mathbf{B}$ and $\mathbf{W}$ sequentially as follows. Let $\mathbf{a} = \mathbf{a}_1$, then it can be shown that (Trendafilov and Vines, 2009; Rencher, 2002; Johnson and Wichern, 2002) $\mathbf{a}$ maximizes the following ratio:

$$\frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}} \tag{2.40}$$

This maximization problem (2.40) is equivalent to the generalized eigenvalue problem given by

$$(\mathbf{B} - \lambda\mathbf{W})\mathbf{a} = \mathbf{0} \quad \Rightarrow (\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I}_p)\mathbf{a} = \mathbf{0}. \tag{2.41}$$

The solution to this equation is the eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$. Consequently, the largest eigenvalue, λ_1, of $\mathbf{W}^{-1}\mathbf{B}$, associated with the eigenvector $\mathbf{a} = \mathbf{a}_1$, is the maximum value of (2.40). The linear combination $\mathbf{a}_1^T\mathbf{x}$ is called the first linear discriminant function. This discriminant function is the most powerful discriminant function. The second powerful discriminant function is given by the linear combination $\mathbf{a}_2^\top\mathbf{x}$, where $\mathbf{a}_2$ maximizes the ratio (2.40) subject to $\text{Cov}(\mathbf{a}_1^\top\mathbf{x}, \mathbf{a}_2^\top\mathbf{x}) = 0$. In general, the k^{th} linear discriminant function is given by the k^{th} linear combination $\mathbf{a}_k^\top\mathbf{x}$ whose coefficient is associated with the k^{th} eigenvector of $\mathbf{W}^{-1}\mathbf{B}$. $\mathbf{a}_k$ maximizes the ratio (2.40) subject to $\text{Cov}(\mathbf{a}_k^\top\mathbf{x}, \mathbf{a}_i^\top\mathbf{x}) = 0$, $i < k$, and $\text{Var}(\mathbf{a}_i^\top\mathbf{x}) = 1, i = 1, \ldots, s$. The power of discrimination of the linear combinations is determined by their eigenvalues associated with their respective vector of coefficients. We consider the eigenvalues to be ranked as $\lambda_1 > \lambda_2 > \cdots > \lambda_s$. The number of (nonzero) eigenvalues s is the rank of $\mathbf{B}$ which is the minimum of $(g-1, p)$. Hence the discriminant function that best separates the group means is $y_1 = \mathbf{a}_1^T\mathbf{x}$.

Subsequently, the remaining discriminant functions ordered in decreasing their power of discrimination are: $y_2 = \mathbf{a}_2^T\mathbf{x}, \ldots, y_s = \mathbf{a}_s^T\mathbf{x}$. From the s eigenvectors, we obtain s discriminant functions (Rencher, 1992, section 8.4). These discriminant functions are uncorrelated, but they are not orthogonal. This is because $\mathbf{W}^{-1}\mathbf{B}$ is not symmetric matrix.

The main objective of Fisher's discriminant analysis is to separate groups. However, it can also be used to classify observations into their respective groups. The assumption of multivariate normality of the g-groups is not necessary to use Fisher's discriminant method. But, the assumption that group covariance matrices are equal and full rank must be fulfilled. That is, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 \cdots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$.

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_s > 0$ denote the $s \leq \min(g-1, p)$ nonzero eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ and let $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_s$ be the corresponding eigenvectors that satisfy $\mathbf{e}^\top\mathbf{W}\mathbf{e} = 1$. Fisher's LDA is obtained by finding a vector of coefficients $\mathbf{a}$ that maximizes (2.40).

The vector that maximizes (2.40) is given by $\mathbf{a}_1 = \mathbf{e}_1$. The linear combination $\mathbf{a}_1^T\mathbf{X}$ is called the first linear discriminant function. This discriminant function is the most powerful discriminant function. The second discriminant function is given by the linear combination $\mathbf{a}_2^\top\mathbf{X}$, where $\mathbf{a}_2 = \mathbf{e}_2$ maximizes the ratio (2.40) subject to $\text{Cov}(\mathbf{a}_1^\top\mathbf{X}, \mathbf{a}_2^\top\mathbf{X}) = 0$. In general, the k^{th} linear discriminant function is given by the k^{th} linear combination $\mathbf{a}_k^\top\mathbf{X}$ whose coefficient is associated with the k^{th} eigenvector of $\mathbf{W}^{-1}\mathbf{B}$. $\mathbf{a}_k = \mathbf{e}_k$ maximizes the ratio (2.40) subject to $\text{Cov}(\mathbf{a}_k^\top\mathbf{X}, \mathbf{a}_i^\top\mathbf{X}) = 0$, $i < k$, and $\text{Var}(\mathbf{a}_i^\top\mathbf{X}) = 1, i = 1, \ldots, s$. The power of discrimination of the linear combinations is determined by the eigenvalues and eigenvec-

tors. Hence, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_s$ give discriminant functions, respectively ranked from highest to lowest degree of discrimination.

Proof. We first convert the maximization problem to one already solved. By the spectral decomposition, $\mathbf{W}$ can be given as $\mathbf{W} = \mathbf{P}^\top \mathbf{\Lambda} \mathbf{P}$ where $\mathbf{P}$ is a matrix whose columns are the normalized eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k$, and $\mathbf{\Lambda}$ is a diagonal matrix given as

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix} \quad \text{with } \lambda_i > 0.$$

Let $\mathbf{\Lambda}^{1/2}$ denote the diagonal matrix with elements $\sqrt{\lambda_i}$. Thus the symmetric square-root matrix $\mathbf{W}^{1/2} = \mathbf{P}^\top \mathbf{\Lambda}^{1/2} \mathbf{P}$ and its inverse $\mathbf{W}^{-1/2} = \mathbf{P}^\top \mathbf{\Lambda}^{-1/2} \mathbf{P}$ satisfy $\mathbf{W}^{1/2}\mathbf{W}^{1/2} = \mathbf{W}$, $\mathbf{W}^{1/2}\mathbf{W}^{-1/2} = \mathbf{I} = \mathbf{W}^{-1/2}\mathbf{W}^{1/2}$ and $\mathbf{W}^{-1/2}\mathbf{W}^{-1/2} = \mathbf{W}^{-1}$. Now, let us set

$$\mathbf{b} = \mathbf{W}^{1/2}\mathbf{a}$$

so $\mathbf{b}^\top \mathbf{b} = \mathbf{a}^\top \mathbf{W}^{1/2}\mathbf{W}^{1/2}\mathbf{a} = \mathbf{a}^\top \mathbf{W}\mathbf{a}$ and $\mathbf{b}^\top \mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}\mathbf{b} = \mathbf{a}^\top \mathbf{W}^{1/2}\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}\mathbf{W}^{1/2}\mathbf{a} = \mathbf{a}^\top \mathbf{B}\mathbf{a}$. Consequently, the maximization problem (2.40) can be reformulated as

$$\max_{\mathbf{b}} \frac{\mathbf{b}^\top \mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}\mathbf{b}}{\mathbf{b}^\top \mathbf{b}}. \tag{2.42}$$

The maximum of this ratio is the largest eigenvalue of $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$, which is λ_1. This maximization occurs when $\mathbf{b} = \mathbf{e}_1$, the normalized eigenvector associated with λ_1. Because $\mathbf{e}_1 = \mathbf{b} = \mathbf{W}^{1/2}\mathbf{a}_1$, or $\mathbf{a}_1 = \mathbf{W}^{-1/2}\mathbf{e}_1$, $\text{Var}(\mathbf{a}_1^\top \mathbf{X}) = \mathbf{a}_1^\top \mathbf{W}\mathbf{a}_1 =$

$\mathbf{e}_1^\top \mathbf{W}^{-1/2}\mathbf{W}\mathbf{W}^{-1/2}\mathbf{e}_1 = \mathbf{e}_1^\top \mathbf{W}^{-1/2}\mathbf{W}^{1/2}\mathbf{W}^{1/2}\mathbf{W}^{-1/2}\mathbf{e}_1 = \mathbf{e}_1^\top \mathbf{e}_1 = 1$. $\mathbf{b} \perp \mathbf{e}_1$ maximizes the preceding ratio when $\mathbf{b} = \mathbf{e}_2$, the normalized eigenvector corresponding to λ_2. For this choice, $\mathbf{a}_2 = \mathbf{W}^{-1/2}\mathbf{e}_2$, and $\text{Cov}(\mathbf{a}_1^\top \mathbf{X}, \mathbf{a}_2^\top \mathbf{X}) = \mathbf{a}_2^\top \mathbf{W}\mathbf{a}_1 = \mathbf{e}_2^\top \mathbf{e}_1 = 0$, since $\mathbf{e}_2 \perp \mathbf{e}_1$. Similarly, $\text{Var}(\mathbf{a}_2^\top \mathbf{X}) = \mathbf{a}_2^\top \mathbf{W}\mathbf{a}_2 = \mathbf{e}_2^\top \mathbf{e}_2 = 1$. We continue in this fashion to determine the remaining discriminant functions. For example, to determine the k^{th} discriminant, we find $\mathbf{b} \perp \mathbf{e}_i$ that maximizes the ratio (2.42), subject to orthogonality constraint, and this is the normalized eigenvector corresponding to λ_k. That is, $\mathbf{b} = \mathbf{e}_k, i < k$. For this choice, the discriminant vector is given as $\mathbf{a}_k = \mathbf{W}^{-1/2}\mathbf{e}_k$, and

$$\mathbf{a}_k^\top \mathbf{W}\mathbf{a}_i = \begin{cases} 1, & i = k, \quad \text{for } i, k = 1, 2, \ldots s \\ 0, & i < k. \end{cases}$$

□

Note that if λ is the eigenvalue of $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$ and $\mathbf{e}$ is its associated eigenvector , then $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}\mathbf{e} = \lambda\mathbf{e}$ and multiplying on the left hand side by $\mathbf{W}^{-1/2}$ gives

$$\mathbf{W}^{-1/2}\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}\mathbf{e} = \lambda\mathbf{W}^{-1/2}\mathbf{e} \quad \text{or} \quad \mathbf{W}^{-1/2}\mathbf{B}(\mathbf{W}^{-1/2}\mathbf{e}) = \lambda(\mathbf{W}^{-1/2}\mathbf{e}).$$

Consequently, $\mathbf{W}^{-1/2}\mathbf{B}$ has the same eigenvalues as $\mathbf{W}^{-1/2}\mathbf{B}\mathbf{W}^{-1/2}$, but the corresponding eigenvector is proportional to $\mathbf{W}^{-1/2}\mathbf{e} = \mathbf{a}$, as shown above.

Let the k^{th} linear discriminant function (LDF) be given by $Y_k = \mathbf{a}_k^\top \mathbf{X}$, where

$$\mathbf{a}_k = \mathbf{W}^{-1/2}\mathbf{b}_k = \begin{bmatrix} a_{k1} \\ a_{k2} \\ \vdots \\ a_{kp} \end{bmatrix}, \text{ and } \mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}, \quad k = 1, 2, \ldots s. \tag{2.43}$$

Then the resulting s linear discriminants $Y_1, Y_2, \ldots, Y_s$ are given as:

$$\begin{aligned} Y_1 &= \mathbf{a}_1^\top \mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots a_{1p}X_p \\ Y_2 &= \mathbf{a}_2^\top \mathbf{X} = a_{21}X_1 + a_{22}X_2 + \cdots a_{2p}X_p \\ \vdots &= \vdots \\ Y_s &= \mathbf{a}_s^\top \mathbf{X} = a_{s1}X_1 + a_{s2}X_2 + \cdots a_{sp}X_p. \end{aligned} \tag{2.44}$$

We can put these functions in matrix form as

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_s \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1^\top \mathbf{X} \\ \mathbf{a}_2^\top \mathbf{X} \\ \vdots \\ \mathbf{a}_s^\top \mathbf{X} \end{bmatrix} = \mathbf{A}^\top \mathbf{X}, \tag{2.45}$$

where $\mathbf{A}$ is a transformation matrix whose k^{th} row is $\mathbf{a}_k^\top$ such that $\mathbf{AWA}^\top = \mathbf{I}_s$. This implies that the components of $\mathbf{Y}$ have unit variances and zero covariances.

The aim of deriving these discriminant functions is to obtain a low-dimensional representation of the data that separates the groups as much as possible. In addition to group separation, the discriminants also give the basis for a classification rule. A reasonable classification rule is one that assigns $\mathbf{y}$ to group k if the square of the distance from $\mathbf{y}$ to $\boldsymbol{\mu}_k$ is smaller than the square of the distance from $\mathbf{y}$ to $\boldsymbol{\mu}_i$ for $i \neq k$.

It is well known that **W** is singular when $p >> n$. Consequently, in the high-dimensional scenario, it is impossible to find the eigenvalues and their associated eigenvectors of $\mathbf{W}^{-1}\mathbf{B}$.

2.4.3 Regression approach to LDA for two groups

Fisher (1936) also used a linear regression approach as an alternative way to derive the linear discriminant function for two groups. The discrimination problem can be viewed as a special case of regression. The components of $\mathbf{x}$ are taken as regressor variables and a dummy variable indicating group membership is taken as a dependent variable. Denote the dependent variable for the i^{th} group on the j^{th} observation by $y_{ij}, j = 1, 2, \ldots, n_i, i = 1, 2$. Then the linear regression between the dependent and the regressor variables is given as

$$y_{ij} = \mathbf{b}^T \mathbf{x}_{ij} + \epsilon_{ij} \tag{2.46}$$

where ϵ_{ij} are error terms. The two values taken by the dependent variable in (2.46) are irrelevant. Fisher, actually, took the values $y_{1j} = \frac{n_2}{n_1+n_2}$ if $\mathbf{x}_{ij} \in \pi_1$ and $y_{2j} = \frac{-n_1}{n_1+n_2}$ if $\mathbf{x}_{ij} \in \pi_2$. The objective is to estimate the parameter $\mathbf{b}$ that best fits the model (2.46). It is estimated by minimizing

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \mathbf{b}^T(\mathbf{x}_{ij} - \bar{\mathbf{x}}))^2$$

where

$$\bar{\mathbf{x}} = \frac{n_1\bar{\mathbf{x}}_1 + n_2\bar{\mathbf{x}}_2}{n_1 + n_2} \tag{2.47}$$

The normal equations are

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(\mathbf{x}_{ij} - \bar{\mathbf{x}})(\mathbf{x}_{ij} - \bar{\mathbf{x}})^T\mathbf{b} = \sum_{i=1}^{2}\sum_{j=1}^{n_i} y_{ij}(\mathbf{x}_{ij} - \bar{\mathbf{x}}) \tag{2.48}$$

Solving (2.48) for **b** gives

$$\mathbf{b} = \mathbf{S^{-1}}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\left[\frac{n_1 n_2(1-c)}{(n_1+n_2)(n_1+n_2-2)}\right] \tag{2.49}$$

where c is a constat. Hence, **b** is proportional to $\mathbf{S^{-1}}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$, the discriminant coefficient (**w**) obtained earlier in (2.22). It is identical with the vector **a** that maximizes (2.38).

Chapter 3

Review of discriminant analysis in high-dimensions

Classical linear discriminant analysis (LDA) does not perform classification effectively when the number of variables, p, is much larger than the number of observations, n, commonly written as $p >> n$. There are two major reasons that classical LDA is not directly applicable in high dimensional settings. First, the sample covariance matrix estimate is singular or nearly singular and cannot be inverted (Guo et al., 2007). This reflects the presence of redundant variables (noise accumulation) that do not significantly contribute to the separation between groups (Qiao et al., 2009). Although we may use the generalized inverse of the covariance matrix, the estimate is highly biased and unstable and will lead to a classifier with poor performance due to the lack of observations. Second, high-dimensionality makes direct matrix operation very difficult if not impossible, hence hindering the applicability of the traditional LDA method.

Several techniques have been developed recently to circumvent the aforementioned problems of LDA in high dimensions. These vary in their assumptions and techniques and their characteristics give four classes as below.

1. **Dimension reduction methods**: these methods involve dimension reduction by setting many parameters to zero. As a result the contribution of the variables associated with those parameters are assumed to be insignificant to the discrimination between classes;

2. **Regularization methods**: these methods regularize the within-class covariance matrix to obtain an invertible covariance matrix. Then, the discriminant vector can be estimated using the classical discrimination methods.

3. **Ratio optimization methods**: these methods focus on the ratio of the between-groups variance to the within-group variance and aim maximize this ratio, perhaps with added constraints to impose sparsity.

4. **Optimal scoring methods**: these methods recast the discriminant analysis problem as a regression problem.

In this chapter, we will briefly review these four classes and then briefly review some miscellaneous methods that do not fit into the classes.

3.1 Dimension reduction Methods

Various discrimination methods use global dimension reduction techniques to circumvent the problems that arise from high dimensionality (Bouveyron et al., 2007). A commonly used method is to first reduce the dimensionality of the data

and then using a classical DA on the dimension-reduced data. This method is called two-stage DA. The process of dimension reduction can be done using different variable selection techniques (Bouveyron et al., 2007) or principal component analysis (PCA) (Jolliffe, 2002). The motivation to use two-stage DA often comes from the context of the application at hand. Fisher LDA can also be used to reduce the dimension for classification purposes. Fisher LDA projects the data on the $(g-1)$ discriminant axes and then classifies the projected data (Bouveyron et al., 2007).

Another perspective on the curse of dimensionality in discriminant analysis is to consider it as an over-parameterized modeling problem. Bouveyron and Brunet-Saumard (2014) argue that a Gaussian model is highly parameterized and that this causes inference problems in high dimensional spaces. It follows that the use of constrained or parsimonious models is a way of avoiding the problem of high-dimensionality in model-based discriminant analysis.

A commonly used way to reduce the number of parameters in a Guassian model is to impose constraints based on assumptions on the parameters of the model. This method can be illustrated by considering an example similar to the constrained Gaussian model given by Bouveyron and Brunet-Saumard (2014). Suppose an unconstrained Gaussian model (the full model) that highly parameterized and contains 20603 parameters when there are $g = 4$ groups and $p = 100$ variables. One possible constraint for reducing the number of parameters is to assume that all groups have the same covariance matrix, i.e. $\mathbf{\Sigma}_i = \mathbf{\Sigma}, \forall_i, i = 1, 2, ..., g$. Note that this model yields Fisher's famous LDA. It is also possible to

assume that the variables are conditionally independent. This assumption implies that each covariance matrix is diagonal, i.e. $\mathbf{\Sigma}_i = \text{diag}(\sigma_{i1}^2, ..., \sigma_{ip}^2)$, where σ_{il}^2 is variance of the l^{th} variable in the i^{th} group. In this case, where groups have the same covariance matrix in addition to the independence assumption, the common covariance matrix will be estimated as, $\mathbf{\Sigma} = \text{diag}(\sigma_1^2, ..., \sigma_p^2)$, where σ_l^2 is variance of the l^{th} variable in each group. Two other constraints are based on the assumption that the covariance matrices are proportional to an identity matrix. They are: when the covariance matrix is spherical in each group $\mathbf{\Sigma}_i = \sigma_i^2 \mathbf{I}_p$, and when it is assumed that the covariance matrices are equal and spherical such that $\mathbf{\Sigma}_i = \Sigma = \sigma^2 \mathbf{I}_p$, for $i = 1, 2..., g$ and $\sigma^2 \in \mathbb{R}$.

For comparison, Table 3.1 lists the most commonly used model assumptions that can be obtained from a Gaussian mixture model with g groups and p variables. The number of parameters can be decomposed into the number of parameters for the proportions $(g-1)$, for the means gp, and for the covariance matrices (last term).

We can see that the full-model is a highly parameterized model. In contrast, the 5^{th} and the 6^{th} models are very parsimonious models (Bouveyron and Brunet-Saumard, 2014). These models, however, work under the strong assumption of independence of variables which may be unrealistic in many discrimination problems. The second model requires estimation of an intermediate number of parameters, in this case 5453. This model is known to be an efficient model in practical classification problems. Furthermore, this model is commonly used when the normality assumption does not hold.

Table 3.1: *Number of parameters to estimate for constrained Gaussian models*

Model	Assumption	No. of parameters	$g = 4$ and $p = 100$
1	Full-Model	$(g-1) + gp + gp(p+1)/2$	20603
2	$\mathbf{\Sigma}_i = \mathbf{\Sigma}$	$(g-1) + gp + p(p+1)/2$	5453
3	$\mathbf{\Sigma}_i = \text{diag}(\sigma_{i1}^2, ..., \sigma_{ip}^2)$	$(g-1) + gp + gp$	803
4	$\mathbf{\Sigma}_i = \mathbf{\Sigma} = \text{diag}(\sigma_1^2, ..., \sigma_p^2)$	$(g-1) + gp + p$	503
5	$\mathbf{\Sigma}_i = \sigma_i^2 \mathbf{I}_p$	$(g-1) + gp + g$	407
6	$\mathbf{\Sigma}_i = \mathbf{\Sigma} = \sigma^2 \mathbf{I}_p$	$(g-1) + gp + 1$	404

3.2 Regularization methods

In this section methods that mainly focus on estimating the within-class covariance matrix using various regularization methods will be briefly reviewed. In general, methods of regularizing the within-class covariance matrix can be categorized into two main groups. The first group of methods make an independent assumption that force the within-class covariance matrix to be diagonal. The second group of methods allow dependence and use various techniques to estimate the full covariance matrix (Clemmensen, 2013).

3.2.1 Independence assumption

Because of the high dimension p and small sample size n, which are often referred to as *large p small n*, estimators of the sample mean and covariance matrix are usually unstable (Wang et al., 2013). Bickel and Levina (2004) have shown that Fisher's LDA is no better than random guessing when $p/n \longrightarrow \infty$. From

the existing literature it is possible to classify the independence rules into two classes (Wang et al., 2013). The first and natural method is to ignore the dependence among the variables, which leads to the so called Naive Bayes Classifier. Some methods that assume independence are given in Tibshirani et al. (2003), Tibshirani et al. (2002), and Dudoit et al. (2002). These methods will be reviewed here briefly, together with a method that involves individual analysis (Fan and Fan, 2008).

3.2.1.1 Nearest shrunken centroids (NSC)

Tibshirani et al. (2003) proposed a method for class prediction in high dimensional microarray studies based on an enhancement of the nearest prototype classifier. This method uses 'shrunken' centroids as prototypes for each class where class centroids are shrunk toward the overall centroid.

In this approach, the covariance matrix is estimated as the diagonal of the full covariance estimate $\hat{\Sigma}_{NSC} = \text{diag}(\hat{\Sigma}) = \text{diag}(s_1^2, s_2^2, ..., s_p^2)$, where s_l^2 $(l = 1, 2, ..., p)$ is variance of the l^{th} variable. Consequently, the group means are shrunk using soft thresholding shrinkage. The absolute value of each $\hat{\Sigma}_{NSC}^{-1}\hat{\mu}_i$ is reduced by an amount Δ and is set to zero if the result is less than zero. That is:

$$\hat{\Sigma}_{NSC}^{-1}\hat{\mu}_i{}^* = \text{sign}(\hat{\Sigma}_{NSC}^{-1}\hat{\mu}_i)(|\hat{\Sigma}_{NSC}^{-1}\hat{\mu}_i| - \Delta)_+ \tag{3.1}$$

where the subscript 'plus' means the positive part ($t_+ = t$ if $t > 0$). If the shrinkage parameter is very large, many of the components (genes) will be eliminated. Hence, Δ tunes the degree of sparsity. In particular, if Δ causes $\hat{\Sigma}_{NSC}^{-1}\hat{\mu}_i$ to shrink to zero for all groups i, then the mean for variable l ,i.e.,$\bar{X}_l$, is the same for all groups. Thus, variable l will not have a contribution to the nearest mean com-

putation. Δ is chosen by cross-validation. Similar methods can also be seen in Dudoit et al. (2002) and Tibshirani et al. (2002).

3.2.1.2 Independence rule (IR)

Bickel and Levina (2004) proposed an independence rule where the covariance matrix is estimated by the diagonal of the covariance matrix. They explained that the 'Naive Bayes' classifier which assumes independence among variables greatly outperforms the Fisher LDA rule under certain conditions of the number of variables grows faster than the number of observations. They considered the problem of discriminating between two groups with p-variate normal distributions $N_p(\mu_1, \mathbf{\Sigma})$ and $N_p(\mu_2, \mathbf{\Sigma})$. A new observation $\mathbf{x}$ is to be assigned to one of these two groups. If μ_1, μ_2, and $\mathbf{\Sigma}$ are known, then the optimal classifier is the Bayes Rule, expressed through the group indicator function $\mathbf{1}$ as:

$$\delta(\mathbf{x}) = \mathbf{1}\{\log\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right) > 0\} = \mathbf{1}(\mu_d^T \mathbf{\Sigma^{-1}}(\mathbf{x} - \mu) > 0) \tag{3.2}$$

where the prior probabilities are assumed to be equal and $\mu_d = \mu_1 - \mu_2$ and $\mu = \frac{\mu_1 + \mu_2}{2}$. Plugging all the parameter estimates directly into the Bayes Rule (2.4) leads to the Fisher rule (FR):

$$\delta_F(\mathbf{x}) = \mathbf{1}(\hat{\mu}_d^T \hat{\mathbf{\Sigma}}^{-1}(\mathbf{x} - \hat{\mu}) > 0). \tag{3.3}$$

Bickel and Levina (2004) assumed that variables are independent, and hence they replaced the off-diagonal elements of $\hat{\mathbf{\Sigma}}$ with zeros. Thus, under this assumption, the covariance matrix is estimated as: $\hat{\mathbf{D}} = \text{diag}(\hat{\mathbf{\Sigma}})$. The resulting discrimination rule is called the independence rule (IR) and is given as

$$\delta_I(\mathbf{x}) = 1(\hat{\mu}_d^T \hat{\mathbf{D}}^{-1}(\mathbf{x} - \hat{\mu}) > 0). \tag{3.4}$$

They compared the performance of Fisher's rule and the independence rule under the worse-case scenario, where $p \longrightarrow \infty$, $n \longrightarrow \infty$, and $p/n \longrightarrow \infty$.

Bickel and Levina (2004) considered two conditions on the properties of $\mathbf{\Sigma}$ and Δ^2. The first condition is given as

$$\frac{\lambda_{\max}(\mathbf{\Sigma})}{\lambda_{\min}(\mathbf{\Sigma})} < \infty.$$

This ratio is called the condition number of $\mathbf{\Sigma}$, where $\lambda_{min}(\mathbf{\Sigma})$ and $\lambda_{max}(\mathbf{\Sigma})$ are the minimum and maximum eigenvalues of $\mathbf{\Sigma}$, respectively. This condition guarantees that both $\mathbf{\Sigma}$ and $\mathbf{\Sigma}^{-1}$ are not ill conditioned. The second condition is $\Delta^2 \geq C^2$, where C is a positive constant and Δ the Mahalanobis distance. This condition ensures that the Mahalanobis distance between the two classes is at least C. Thus C is a measure of the difficulty of the classification. The larger the value of C, the easier the classification is.

For fixed p, the worst-case misclassification rate of $\delta_F(\mathbf{x})$, denoted by $W(\delta_F(\mathbf{x}))$, converges asymptotically to the optimal Bayes risk $(1-\phi(C/2))$. That is, $W(\delta_F(\mathbf{x}))$ $\longrightarrow (1-\phi(C/2))$, while the misclassification rate of $\delta_I(\mathbf{x})$ converges to something strictly greater than the Bayes risk. Hence, $\delta_F(\mathbf{x})$ is asymptotically optimal for low dimensional problems. However, in high a dimensional setting, i.e., when $p > n$, $\delta_F(\mathbf{x})$ is not asymptotically optimal because $\hat{\mathbf{\Sigma}}^{-1}$ is ill-conditioned. Taking the Moore-Penrose generalized inverse ($\hat{\mathbf{\Sigma}}^{-}$) in place of $\hat{\mathbf{\Sigma}}^{-1}$ in (3.3), and assuming $n_1 = n_2$, Bickel and Levina (2004) have shown that under some regularity conditions, if $p/n \longrightarrow \infty$, then $\delta_F(\mathbf{x}) \longrightarrow 1/2$. This suggests that Fisher's LDA performs asymptotically no better than random guessing when $p >> n$. This poor performance of Fisher's LDA is due to the diverging spectra charac-

teristic of high-dimensional covariance matrices. This is the difficulty of high dimensional classification using the classical methods. Consequently, Bickel and Levina (2004) took the diagonal estimate of the covariance matrix for classification purposes. They derived the relative efficiency of IR to the FR theoretically and they concluded that the IR performs much better than the Fisher rule when $p >> n$.

3.2.1.3 Features annealed independence rules (FAIR)

Fan and Fan (2008) studied the impact of high dimensionality on classification. They identified that the difficulty of high dimensional classification is essentially caused by the existence of many noise features that do not contribute to the reduction of classification error. For example, if we need to estimate the class mean vectors and covariance matrix for the Fisher's discriminant rule, each parameter can be estimated correctly. However, aggregated estimation error over many variables can be very large and this significantly causes to increase the misclassification rate.

Fan and Fan (2008) explained that when there are only few variables that account for most of the variation in the data, taking all variables will increase the misclassification error. They demonstrated that even for the independence classification rule, classification using all the features (variables) can be as poor as random guessing due to noise accumulation in estimating population means in high dimensional setting. Furthermore, they demonstrated that almost all linear discriminants can perform as poorly as random guessing in such situations. As a result, they proposed a method that selects a subset of the variables before

the main classification is performed. This method selects the statistically significant variables using two-sample t-statistics , and then the Independence rule is applied to this set of variables. Fan and Fan (2008) called the resulting method as Feature Annealed Independence Rule (FAIR). They used the upper bound of classification error to select the optimal number of variables.

Fan and Fan (2008) compared the performance of their classifier (i.e., FAIR) with the independence rule without variable selection and with a version of FAIR called oracle assisted FAIR. Oracle assisted FAIR addresses an ideal situation in which the important variables are located at the first m coordinates and the variable selection task is to merely select m to minimize the misclassification error. This assumes there is perfect information about the relative importance of the different variables.

Another group of methods uses projection for dimension reduction. Most of the the commonly used projection methods have been widely applied to classification problems involving gene expression data (Fan and Fan, 2008). These projection methods find directions by giving much more weight to variables that have large classification power. However, Fan and Fan (2008) explained that linear projection methods are likely to perform poorly unless the projection vector is sparse, i.e., when the effective number of variables is small. This is because of the noise accumulation that is seen in high dimensional problems.

There is a huge literature on classification. In high dimensional classification minimizing classification error is given much more concern than the accuracy of the estimated parameters. Hence, estimating all covariance matrix and the

class mean vectors will result in very high accumulation errors and thus high classification error.

3.2.1.4 Penalized linear discriminant analysis (PLDA)

Witten and Tibshirani (2011) have proposed a penalized LDA method to achieve interpretability in high dimensional setting. This method penalizes the discriminant vectors in Fisher's discriminant problem. The resulting discriminant problem is not convex, so they use a minorization-maximization method to optimize it efficiently under convex penalties that are applied to the discriminant vectors. In particular, this method uses L_1 and fused lasso penalties. The method is equivalent to recasting Fisher's discriminant problem as a biconvex problem.

It is known that Fisher's discriminant problem finds a low dimensional projection of the observations such that the between-class variance is large relative to the within-class variance, i.e. it sequentially solves

$$\max_{\mathbf{a}_k}(\mathbf{a}_k^T\hat{\mathbf{\Sigma}}_b\mathbf{a}_k) \text{ subject to } \mathbf{a}^{\mathbf{T}}{}_k\hat{\mathbf{\Sigma}}_w\mathbf{a}_k \leq 1,\ \mathbf{a}_k^T\hat{\mathbf{\Sigma}}_w\mathbf{a}_i = 0,\ \forall i < k \tag{3.5}$$

where $\hat{\mathbf{\Sigma}}_b$ and $\hat{\mathbf{\Sigma}}_w$ are the sample estimates of the between-classes and within-class covariance matrices, respectively, and variables have been centered to have mean 0. The solution to problem (3.5) gives $\mathbf{a}_k$ as the k^{th} discriminant vector $(k = 1, 2, ..., g-1)$. This discrimination problem is generally written with the inequality constraint. An equality constraint is taken if $\hat{\mathbf{\Sigma}}_w$ has full rank.

In this discrimination framework, a classification rule is obtained by computing $\mathbf{X}\hat{\mathbf{a}}_1, ..., \mathbf{X}\hat{\mathbf{a}}_{g-1}$ and assigning each observation to its nearest centroid in the transformed space. Alternatively, it is possible to transform the observations by using only the first $s < g-1$ discriminant vectors to perform reduced rank clas-

sification (Witten and Tibshirani, 2011). Problem (3.5) can be solved by substituting $\tilde{\mathbf{a}}_k = \hat{\boldsymbol{\Sigma}}_w^{1/2}\mathbf{a}_k$, where $\hat{\boldsymbol{\Sigma}}_w^{1/2}$ is the symmetric matrix square root of $\hat{\boldsymbol{\Sigma}}_w$. Hence, Fisher's discrimination problem is reduced to standard eigen problem.

Various methods have been proposed to modify problem (3.5) to tackle the singularity problem. For example Krzanowski et al. (1995) modified problem (3.5) to find a unit vector $\mathbf{a}$ that maximizes the objective function subject to $\mathbf{a}_k^T\hat{\boldsymbol{\Sigma}}_w\mathbf{a}_k = 0$; others have used a positive definite estimate of $\boldsymbol{\Sigma}_w$.

Witten and Tibshirani (2011) took the diagonal estimate of the within-class covariance matrix to solve problem (3.5). Hence, they rewrite problem (3.5) as

$$\max_{\mathbf{a}_k}(\mathbf{a}_k^T\hat{\boldsymbol{\Sigma}}_b\mathbf{a}_k) \text{ subject to } \mathbf{a}^{\mathbf{T}}{}_k\hat{\mathbf{D}}\mathbf{a}_k \leq 1,\ \mathbf{a}_k^T\hat{\mathbf{D}}\mathbf{a}_i = 0,\ \forall_i < k \tag{3.6}$$

where $\hat{\mathbf{D}} = \text{diag}(\hat{\boldsymbol{\Sigma}}_w)$. Hence, they used a minorization-maximization algorithm to solve (3.6).

Witten and Tibshirani (2011) further modified problem (3.6) by including a convex penalty function P_k on $\mathbf{a}_k$. The maximization problem becomes

$$\max_{\mathbf{a}_k}(\mathbf{a}_k^T\hat{\boldsymbol{\Sigma}}_b\mathbf{a}_k - P_k(\mathbf{a}_k)) \text{ subject to } \mathbf{a}_k^T\hat{\mathbf{D}}\mathbf{a}_k \leq 1,\ \text{and } \mathbf{a}_k^T\hat{\mathbf{D}}\mathbf{a}_i = 0,\ \forall_i < k. \tag{3.7}$$

When $k = 1$, the first penalized discriminant vector $\mathbf{a}_1$ will be the solution to the problem

$$\max_{\mathbf{a}_1}(\mathbf{a}_1^T\hat{\boldsymbol{\Sigma}}_b\mathbf{a}_1 - P_1(\mathbf{a}_1)) \text{ subject to } \mathbf{a}_1^T\hat{\mathbf{D}}\mathbf{a}_1 \leq 1. \tag{3.8}$$

Problem (3.8) is closely related to penalized PCA, as described for example in Jolliffe et al. (2003). In fact, (3.8) would be exactly penalized PCA if $\hat{\mathbf{D}} = \mathbf{I}$, where $\mathbf{I}$ is an identity matrix. Witten and Tibshirani (2011) considered two specific forms for P_k, the L_1-penalty and the fused lasso penalty to solve problem (3.7). The

fused lasso penalty (Tibshirani et al., 2005) requires ordering of the variables is known apriori. It achieves sparsity by solving

$$\max_{\mathbf{a}_k} \left(\mathbf{a}_k^T \hat{\mathbf{\Sigma}}_b \mathbf{a}_k - \lambda \sum_{l=1}^{p} |\hat{\sigma}_l a_{kl}| \right) \text{ subject to } \mathbf{a}_k^T \hat{\mathbf{D}} \mathbf{a}_k \leq 1, \text{ and } \mathbf{a}_k^T \hat{\mathbf{D}} \mathbf{a}_i = 0, \ \forall_i < k \tag{3.9}$$

where $\hat{\sigma}_l$ is the within-class standard deviation for the l^{th} variable. λ controls the degree of sparsity. When the tuning parameter λ is large, some elements of the solution **a** will be exactly equal to 0. Hence, the resulting discriminant vectors will be sparse.

A visible drawback the penalized LDA is that it only uses the diagonal elements of the covariance matrix. The correlated variables could have any effect on the discrimination. Furthermore, a criticism of this method is that little is known about the theoretical properties of the estimator in (3.9) (Mai et al., 2012).

In general, most of the independence rules of classification assume that all groups have equal covariance matrices and variables are independent. As a result, they use the diagonal covariance matrix of the common covariance matrix. The assumption of equal covariance matrices and independence is explained by Model 4 in Table 3.1. We can see from Table 3.1 that the 4^{th} model is very parsimonious model that has only 503 parameters to be estimated whereas the full-model has 20603 parameters to be estimated when $g = 4$ and $p = 100$. If we consider high-dimensionality in discriminant analysis as an over-parameterized modeling problem, the independence rules are effective in dimension reduction. In this perspective, the discrimination methods that assume independence are typical examples of dimension reduction methods.

3.2.2 Dependence assumption

The independence rule assumes that there is no correlation between variables in the high dimensional setting, and hence $\boldsymbol{\Sigma}$ is diagonal. However, in most microarray studies, correlation between different genes is an inevitable characteristic of the data. For example, Wu et al. (2009) pointed out that there is often a group of correlated genes in gene expression studies in which correlations cannot be ignored and the covariance information can help to minimize misclassification rate. Fan et al. (2012) and Mai et al. (2012) found that the independence rule leads to inefficient variable selection and inferior classification. They also showed that optimal classification error by using the independence rule increases as correlation between variables increases, when $\rho \in [0, 1)$. Where ρ is the coefficient of correlation between variables. However, Fan et al. (2012) have not explained the effect of correlation on classification when $\rho \notin [0, 1)$. Mardia et al. (1979) examine the general effect of correlation between variables on classification.

For illustration, consider two bivariate normal populations. Suppose the covariance matrix between the two variables is given as

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

and let $\boldsymbol{\mu}_d$ be the difference between the means of the two normal populations, so $\boldsymbol{\mu}_d = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. As the population distributions are bivariate, we can put $\boldsymbol{\mu}_d = (\mu_{d1}, \mu_{d2})^T$. Then, the Mahalanobis distance between the two groups is

$$\Delta^2 = \boldsymbol{\mu}_d^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_d = \frac{1}{1-\rho^2}(\mu_{d1}^2 + \mu_{d2}^2 - 2\rho\mu_{d1}\mu_{d2}),$$

and, if the variables are uncorrelated,

$$\Delta_0^2 = \mu_{d1}^2 + \mu_{d2}^2,$$

where Δ_0^2 denotes the Mahalanobis distance with uncorrelated variables. Thus the correlation will reduce the misclassification rate (i.e. improve discrimination) if and only if $\Delta^2 > \Delta_0^2$. This occurs when

$$\rho[(1+h^2)\rho - 2h] > 0 \ \text{, where } \ h = \mu_{d2}/\mu_{d1}.$$

This simple example shows that the misclassification rate will be reduced if ρ does not lie between 0 and $2h/(1+h^2)$, while a very small value of ρ can actually cause poor classification. Note that any positive correlation between variables can increase the misclassification rate if $\mu_{d1} = \mu_{d2}$ (Mardia et al., 1979, Section 11.8). Therefore, the independence rule, in general, is not an efficient method for classification when $\rho \in [0, 1)$.

The other approach to regularization of the estimate of the within-class covariance matrix is to take into account the dependence between variables.

When $p > n$, the sample covariance matrix is singular. Although the inverse of the within-class covariance matrix may be estimated by the generalized inverse, the estimate will be very unstable due to lack of observations (Ramey and Young, 2013; Guo et al., 2007). This instability can be examined through the spectral decomposition of $\hat{\mathbf{\Sigma}}^-$, where

$$\hat{\mathbf{\Sigma}}^- = \sum_{l=1}^{p} \frac{\mathbf{v}_l \mathbf{v}_l^T}{e_l},$$

e_l is the l^{th} largest eigenvalue of $\hat{\mathbf{\Sigma}}$ and $\mathbf{v}_l$ is the associated eigenvector. It is known that the estimated eigenvalues of $\hat{\mathbf{\Sigma}}$ are biased, with smaller eigenvalues

being underestimated (Seber, 2004), and the bias increases as the total number of observations decreases relative to the number of variables. As pointed out in Ramey and Young (2013), the smallest eigenvalues and the directions associated with their corresponding eigenvectors highly influence the estimator of $\boldsymbol{\Sigma}^{-1}$, causing classical LDA to produce an unstable and unreliable classification rule when $p >> n$.

Various regularization techniques have been proposed to correct for the instability of $\hat{\boldsymbol{\Sigma}}_w^-$. Some these focus on regularizing the covariance matrix using a shrinkage estimation method. The shrinkage estimation shrinks the extreme eigenvalues of $\hat{\boldsymbol{\Sigma}}_w$ toward more moderate values and, thus, more stable values. That is, the shrinkage method simultaneously decreases larger eigenvalues and increases smaller eigenvalues, reducing bias (Ramey and Young, 2013; Clemmensen, 2013). One approach to regularizing a covariance matrix is to augment it with a matrix that is proportional to identity matrix. In this section, we review four methods that adopt this approach: regularized discriminant analysis (RDA), penalized discriminant analysis (PDA), regularized linear discriminant analysis (RLDA), sparse linear discriminant analysis methods (sLDA).

3.2.2.1 Regularized discriminant analysis (RDA)

Friedman (1989) has proposed a regularized discriminant analysis in small sample high dimensional classification problems with g groups, where the covariance matrices are not assumed to be equal. Friedman (1989) has estimated the k^{th} class covariance matrix using the following regularization:

$$\hat{\boldsymbol{\Sigma}}_k(\lambda, \gamma) = (1-\gamma)\hat{\boldsymbol{\Sigma}}_k(\lambda) + \gamma\Big[\frac{\text{trace}(\hat{\boldsymbol{\Sigma}}_k(\lambda))}{p}\Big] I_p, \tag{3.10}$$

where

$$\hat{\boldsymbol{\Sigma}}_k(\lambda) = \frac{(1-\lambda)(n_k-1)\hat{\boldsymbol{\Sigma}}_k + \lambda(n-g)\hat{\boldsymbol{\Sigma}}}{(1-\lambda)(n_k-1) + \lambda(n-g)},$$

and the parameters, $0 \leq \lambda \leq 1$ and $0 \leq \gamma \leq 1$ are chosen to minimize the misclassification risk. λ controls the contribution of $\hat{\boldsymbol{\Sigma}}_k$ towards $\hat{\boldsymbol{\Sigma}}$, and the regularization parameter γ controls the shrinkage (ridge) of $\hat{\boldsymbol{\Sigma}}_k(\lambda)$ toward a multiple of the identity matrix. As noted above, this shrinkage has the effect of decreasing the larger eigenvalues and increasing the smaller ones.

The discrimination rule for g-groups with unequal covariance matrices is to assign $\mathbf{x}$ to group k if $d_k^Q(\mathbf{x}) = \max_{1\leq i\leq g} d_i^Q(\mathbf{x})$, where $d_i^Q(\mathbf{x})$ is the quadratic discriminant score for the i^{th} group $(i = 1, 2, ..., g)$ given by

$$d_i^Q(\mathbf{x}) = -\frac{1}{2}\ln|\hat{\boldsymbol{\Sigma}}_i| - \frac{1}{2}(\mathbf{x}-\hat{\mu_i})^T\hat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{x}-\hat{\mu_i}) + \ln p_i. \tag{3.11}$$

Johnson and Wichern (2002) and Friedman (1989) used $\hat{\boldsymbol{\Sigma}}_i^{-1}(\lambda,\gamma)$ in place of $\hat{\boldsymbol{\Sigma}}_i^{-1}$ in (3.11) when the number of variables is very large relative to the number of observations.

It can be observed from (3.10) that for $\lambda = 1$, $\hat{\boldsymbol{\Sigma}}_k(\lambda)$ reduces to $\hat{\boldsymbol{\Sigma}}$. This parameter controls the shift between linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). In LDA the decision surface is linear, while the decision boundary in QDA is nonlinear. Regularized discriminant analysis shrinks the separate covariances of QDA toward a common covariance as in LDA. As a result, RDA is an intermediate between LDA and QDA (Friedman, 1989).

An Attractive feature of this regularization approach is that it identifies and uses LDA and QDA at different settings. However, although this regularization method requires the variance of the parameter estimates, it is associated with

increased bias. That is, there is a trade-off between bias and variance. Moreover, this approach does not incorporate the idea of sparsity.

3.2.2.2 Penalized discriminant analysis (PDA)

Hastie et al. (1995) developed a penalized discriminant analysis based on the optimal scoring approach. In this case, the regularization of the within-class covariance matrix is given by

$$\tilde{\mathbf{\Sigma}}_w = \hat{\mathbf{\Sigma}}_w + \gamma I_p, \tag{3.12}$$

where the parameter $\gamma \geq 0$, controls the degree of diagonalization of the within-class covariance matrix. That is, taking $\gamma = 0$ leads to the estimation of the full covariance matrix, and taking $\gamma \longrightarrow \infty$ results in an identity matrix as the estimate of the covariance matrix (Clemmensen, 2013; Hastie et al., 1995). Furthermore, Hastie et al. (1995) have proposed a more general regularization:

$$\tilde{\mathbf{\Sigma}}_w = \hat{\mathbf{\Sigma}}_w + \gamma \Omega, \tag{3.13}$$

where Ω is a $p \times p$ regularization matrix. This penalization differs from the previous one in (3.12) by the fact that it also penalizes the correlations between the predictors.

A limitation of this method is that it does not include any sparsity technique to select a small number of variables. Hence, this method cannot provide us results that will will be easily interpretable in high-dimensional settings.

3.2.2.3 Regularized linear discriminant analysis (RLDA)

Guo et al. (2007) introduced a covariance regularization technique that is closely related to the method used in PDA. In this case, the within-class covariance ma-

trix is estimated as

$$\tilde{\Sigma}_w = \alpha\hat{\Sigma}_w + (1-\alpha)I_p, \tag{3.14}$$

where $\alpha \in [0,1]$. Taking $\alpha = 0$ gives a diagonal estimate of the within-class covariance matrix, and taking $\alpha = 1$ gives a full estimate of $\hat{\Sigma}_w$. It is known that $\hat{\Sigma}_w$ is an estimate of the correlation matrix if the data is normalized. In this situation the RDA is equivalent to the correlation matrix estimate in PDA. Hence, Guo et al. (2007) used the inverse of $\tilde{\Sigma}_w$ instead of $\hat{\Sigma}^{-1}$ to predict group-membership of observations when the class prior probabilities are assumed equal.

This method introduces sparsity by shrinking the class means, putting

$$\tilde{\Sigma}_w^{-1}\hat{\mu}_i^{*} = \text{sign}(\tilde{\Sigma}_w^{-1}\hat{\mu}_i)(|\tilde{\Sigma}_w^{-1}\hat{\mu}_i| - \Delta)_+. \tag{3.15}$$

This is similar to NSC, but with a different $\tilde{\Sigma}$, where Δ is a positive constant that controls the degree of sparsity. Variable selection using this form of shrunken centroid is, in general, considered conservative because it includes a large number of variables (Clemmensen, 2013). Hence, this type of variable selection does not achieve the required sparsity in high-dimensional discriminant analysis. In deed, even though variable selection and dimension reduction are almost essential in high dimensional discriminant analysis, most of the methods mentioned above focus solely on regularizing the covariance matrix, with the aim of tackling the singularity problem. Hence, their most obvious limitation is that they give less attention to sparsity.

3.2.2.4 Sparse LDA (sLDA) for testing gene pathway

Wu et al. (2009) developed a unified framework to jointly test the significance of a pathway and to select a subset of genes that drive the significant pathway

effect. They decompose each gene pathway into a single score by using a regularized form of LDA to achieve dimension reduction and gene selection (sparsity). They considered two-group sparse LDA in high-dimensions. LDA estimates the discriminant direction **a** by maximizing the ratio of between-class variance to the within-class variance, i.e the generalized Rayleigh quotient:

$$\hat{\mathbf{a}} = \max_{\mathbf{a}} \frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}}. \tag{3.16}$$

This finds **a** by solving (3.16) subject to an additional L_1 constraint on **a**. Applying an L_1 constraint ensures that some a_l will be estimated as exactly zero and the corresponding variables will not contribute to the discrimination direction. Moreover, Wu et al. (2009) noted that in the two-class setting, the rank of $\mathbf{B}$ is 1. Hence, (3.16) can be reformulated as

$$\hat{\mathbf{a}} = \min_{\mathbf{a}} \ \mathbf{a}^T\mathbf{W}\mathbf{a} \ \text{ subject to } \ \hat{\mu}_d^T\mathbf{a} = 1 \ , \sum_{l=1}^{p} |a_l| \leq \tau. \tag{3.17}$$

The value of τ controls the degree of sparsity. When τ is small, some of the a_l will be exactly zero. In general, τ may be selected by maximizing the cross-validated (CV) quotient in (3.16) (Wu et al., 2009).

Zou and Hastie (2005) showed that, in the linear regression setting, addition of an L_2 (penalty) improves prediction and variable selection in cases where predictors are highly correlated. In the same manner, Wu et al. (2009) used the L_2 penalty to regularize the within-class covariance matrix. As a result, they used the regularized within-class covariance matrix, $\mathbf{W}$ in (3.17) , which is similar to the regularization applied by Hastie et al. (1995). That is, $\mathbf{W}$ is replaced by $\tilde{\mathbf{W}}$

where

$$\tilde{\mathbf{W}} = \mathbf{W} + \gamma I_p,$$

I_p is the $p \times p$ identity matrix, and γ is a shrinkage parameter used to stabilize the covariance matrix. Wu et al. (2009) took $\gamma = 2\log(p)/n$ when they applied the sparse LDA to pathway testing. If a large value of γ is applied, then the regularized within-class covariance matrix essentially mimics the identity matrix and the procedure approaches the shrunken centroid method (Wu et al., 2009; Guo et al., 2007).

Even though this method has performed well for tasks including gene pathway identification and gene selection, it is not clear whether it works effectively in general problems of discrimination. Furthermore, as pointed out by Mai et al. (2012), the theoretical properties of discrimination are not clearly addressed. It is also limited to discrimination between only two groups.

3.3 Ratio optimization methods

In classical LDA, the linear combination $\mathbf{Y} = \mathbf{XA}$ is a linear transformation of the original data $\mathbf{X}$ into a lower dimensional vector space $\mathbf{Y}$. The goal of Fisher's LDA is to find a $p \times s$ ($s < p$) transformation matrix $\mathbf{A}$ that produces maximum separation between groups by maximizing the ratio of the between groups covariance matrix (B) relative to the within-groups covariance matrix ($\mathbf{W}$). Note that the transformation matrix (orientation) $\mathbf{A}$ is a $p \times s$ rectangular matrix given as $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_s)$, where $\mathbf{a}_i, i = 1, 2, ..., s$ is a column vector of the orientation $\mathbf{A}$. The optimal $\mathbf{A}$ maximizes the Fisher's criterion function $(f(\mathbf{A}))$ (Sharma and

Paliwal, 2008), which is given as

$$f(\mathbf{A}) = \frac{|\mathbf{A}^T\mathbf{B}\mathbf{A}|}{|\mathbf{A}^T\mathbf{W}\mathbf{A}|} \tag{3.18}$$

where $|.|$ is the determinant. Suppose $\mathbf{a}$ is the first column of $\mathbf{A}$, then the standard discrimination problem is given as,

$$\max_{\mathbf{a}} \frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}}. \tag{3.19}$$

Simplifying (3.19) gives that $\mathbf{a}$ is the solution of the conventional eigenvalue problem,

$$\mathbf{W}^{-1}\mathbf{B}\mathbf{a} = \lambda\mathbf{a}. \tag{3.20}$$

That is, $\mathbf{a}$ is an eigenvector that corresponds to the largest eigenvalue (λ). It can be observed from (3.19) that the explicit solution of the orientation can be found when $\mathbf{W}$ is non-singular. However, it is not possible to find the orientation $\mathbf{A}$ by using (3.19) when $\mathbf{W}$ is singular. To overcome this problem, many methods have been proposed including application of intermediate techniques like PCA prior to the application of LDA. The PCA technique is used in such a way that the projected vectors on s-dimensional space give a full rank $\mathbf{W}$. Thereby the computation of the inverse of $\mathbf{W}$ is feasible and thus $\mathbf{A}$ can then be found by the basic LDA. However, the application of intermediate techniques sacrifices some classification performance (Sharma and Paliwal, 2008). Here we review methods that aim to choose $\mathbf{A}$ to maximize (3.18) or $\mathbf{a}$ to maximize (3.19).

3.3.1 A gradient LDA

Sharma and Paliwal (2008) addressed the task of finding the orientation $\mathbf{A}$ that maximizes the function $f(\mathbf{A})$ in (3.18). They proposed a direct computation

of $\mathbf{A}$ by applying a gradient descent method on Fisher's criterion function.

The reciprocal of Fisher's criterion can be denoted as $\hat{J}(\mathbf{A}) = 1/f(\mathbf{A})$, and then the maximization problem becomes a minimization problem, where the goal is now to find the orientation $\mathbf{A}$ that minimizes $\hat{J}(\mathbf{A})$. They derived the gradient LDA method by finding the derivative of $\hat{J}(\mathbf{A})$. Then $\mathbf{A}$ is updated using gradient descent method while normalizing the column vectors of $\mathbf{A}$ in each iteration. They have shown that the derivative of $\hat{J}(\mathbf{A})$ is:

$$\frac{\partial \hat{J}(\mathbf{A})}{\partial \mathbf{A}} = 2\hat{J}(\mathbf{A})\big[\mathbf{W}\mathbf{A}(\mathbf{A}^T\mathbf{W}\mathbf{A})^{-1} - \mathbf{B}\mathbf{A}(\mathbf{A}^T\mathbf{B}\mathbf{A})^{-1}\big]. \tag{3.21}$$

We observed from (3.20) that the $p \times p$ within-groups covariance matrix $\mathbf{W}$ is not invertible when $p > n$. However, $(\mathbf{A}^T\mathbf{W}\mathbf{A})$ and $(\mathbf{A}^T\mathbf{B}\mathbf{A})$ are full rank $s \times s$ matrices, so their inverse can be computed to find the derivative of $\hat{J}(\mathbf{A})$ in (3.21). Therefore, the gradient descent algorithm can be used to solve for the values of $\mathbf{A}$ by normalizing each of the column vectors of $\mathbf{A}$ separately with $\hat{J}(\mathbf{A})$ updated iteratively. The iterative process of the algorithm can be terminated when $J(\mathbf{A})$ becomes stable.

The good side of gradient LDA proposed by Sharma and Paliwal (2008) is that it is based on a direct approach to LDA and preserves the basic information for classification. However, the method does not incorporate any sparsity procedure. Consequently, interpretation of the discriminant function is difficult if this method is directly employed in high dimensional discriminant analysis.

3.3.2 Variable selection in discriminant analysis via the Lasso

Trendafilov and Jolliffe (2007) proposed a procedure for variable selection using the lasso in discriminant analysis. The lasso approach is applied to improve

the interpretability of the canonical variables. They modified the LDA in a PCA fashion to get orthogonal projections of the original data space that maximize the discrimination between groups. To achieve this, they formulate the LDA problem in (3.19) as

$$\max_{\mathbf{a}_i} \frac{\mathbf{a}_i^T \mathbf{B} \mathbf{a}_i}{\mathbf{a}_i^T \mathbf{W} \mathbf{a}_i} \text{ subject to } \mathbf{a}_i^T \mathbf{W} \mathbf{a}_i = 1, \text{ and } \mathbf{a}_i^T \mathbf{W} \mathbf{a}_j = 0, \text{ for } i \neq j. \tag{3.22}$$

Then they reformulate the LDA objective function in (3.22) subject to PCA constraints:

$$\max_{\mathbf{a}_i} \frac{\mathbf{a}_i^T \mathbf{B} \mathbf{a}_i}{\mathbf{a}_i^T \mathbf{W} \mathbf{a}_i} \text{ subject to } \mathbf{a}_i^T \mathbf{a}_i = 1, \text{ and } \mathbf{a}_i^T \mathbf{A}_i = \mathbf{0}_{i-1}^T, \tag{3.23}$$

where $\mathbf{A}_{i-1}$ is a $p \times (i-1)$ matrix defined as $\mathbf{A}_{i-1} = (\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_{i-1})$. The solutions $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_s)$ are called orthogonal canonical variates.

Trendafilov and Jolliffe (2007) assumed that the within-group covariance matrix is non-singular. Consequently, to find the standard canonical variates, they were able to use the Cholesky factorization of $\mathbf{W}$, i.e. $\mathbf{W} = \mathbf{U}^T\mathbf{U}$ in (3.22), where $\mathbf{U}$ is the positive-definite upper-triangular matrix. Furthermore, to achieve more easily interpretable canonical variates, they included additional lasso constraints. Specifically, they defined the PCA-like LDA problems as:

$$\max_{\mathbf{a}} \mathbf{a}^T \mathbf{U}^{-T} \mathbf{B} \mathbf{U}^{-1} \mathbf{a}$$

and

$$\max_{\mathbf{a}} \frac{\mathbf{a}^T \mathbf{B} \mathbf{a}}{\mathbf{a}^T \mathbf{W} \mathbf{a}},$$

both subject to $||\mathbf{a}||_1 \leq t$, $||\mathbf{a}||_2^2 = 1$ and $\mathbf{a}_i^T \mathbf{A}_i = \mathbf{0}_{i-1}^T$. Trendafilov and Jolliffe (2007) introduced an external penalty function P so as to eliminate the Lasso inequality constraint. The idea is to penalize a unit vector $\mathbf{a}$ which does not satisfy the

LASSO constraint by reducing the value of the new objective function. Thus, the LDA problems are modified as follows:

$$\max_{\mathbf{a}} \left[\mathbf{a}^T\mathbf{U}^{-T}\mathbf{B}\mathbf{U}^{-1}\mathbf{a} - \mu P(||\mathbf{a}||_1 - t)\right] \tag{3.24}$$

and

$$\max_{\mathbf{a}} \left[\frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}} - \mu P(||\mathbf{a}||_1 - t)\right], \tag{3.25}$$

both subject to $||\mathbf{a}||_2^2 = 1$ and $\mathbf{a}_i^T\mathbf{A}_i = \mathbf{0}_{i-1}^T$.

The penalty function P is zero if the Lasso constraint is fulfilled. It switches on the penalty μ (a large positive number) if the Lasso constraint is violated. Moreover the more severe violations are penalized more heavily. A typical example of an exterior penalty function for inequality constraints is the Zangwill penalty function $P(x) = \max(0, x)$, which was used in this method.

Finally, they employed a gradient method to solve the PCA-like problems in (3.24) and (3.25). However, the penalty function P and the Lasso constraint are not differentiable and thus the gradient cannot be computed directly. To overcome this problem the following smoothing (Trendafilov and Jolliffe, 2007)) is used:

$||\mathbf{a}||_1 = \mathbf{a}^T\text{sign}(\mathbf{a}) \approx \mathbf{a}^T\tanh(\gamma\mathbf{a})$ and $P(x) = \max(0, x) \approx \frac{x(1+\tanh(\gamma x)}{2}$, for some large γ, e.g. $\gamma = 1000$. Let, the functions in (3.24) and (3.25) be denoted by $F_\mu(\mathbf{a})$:

$$F_\mu(\mathbf{a}) = \mathbf{a}^T\mathbf{U}^{-T}\mathbf{B}\mathbf{U}^{-1}\mathbf{a} - \mu P(\mathbf{a}^T\tanh(\gamma\mathbf{a}) - t) \tag{3.26}$$

and

$$F_\mu(\mathbf{a}) = \frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}} - \mu P(\mathbf{a}^T\tanh(\gamma\mathbf{a}) - t). \tag{3.27}$$

The loadings of the canonical variates $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_s$ can be computed as solutions of s initial value problems for the following ordinary differential equations:

$$\frac{d\mathbf{a}_i}{dt} = \Pi_i \bigtriangledown_{F_\mu(\mathbf{a}_i)} (\mathbf{a}_i), \tag{3.28}$$

starting with an initial value $\mathbf{a}_{i,in}$ with $||\mathbf{a}_{i,in}||_2^2 = 1$ for $i = 1, 2, ..., s$.

The good point of this method is that it makes interpretation simple in linear discriminant analysis. However, since it assumes that the within-covariance matrix is not ill-conditioned, this method is limited to the low dimensional LDA. But, as concluded by Trendafilov and Jolliffe (2007), the method can be easily applied to high-dimensional problems using some data pre-processing procedure.

3.3.3 A sparse LDA algorithm based on subspaces

Ng et al. (2011) presented a sparse LDA algorithm for high-dimensional objects in subspaces. They noted that, in high dimensional data, groups of observations often exist in subspaces rather than in the entire space. That is, each group is a set of observations identified by a subset of dimensions and different groups are represented in different subsets of dimensions. For this setup, Ng et al. (2011) proposed an algorithm called the gradient flow method on the orthogonal constraint. This method helps to find an explicit solution, but it does not correspond to classical LDA.

The gradient flow algorithm considers that different dimensions make different contributions (i.e. weights) to the identification of objects in a group. Consequently, this method tries to find a sparse LDA by simultaneously maximizing the ratio of the between groups covariance matrix to the within-groups covariance matrix while minimizing the weight sparsity of discriminant vectors. As a

result, Ng et al. (2011) formulate the following optimization problem when $\mathbf{W}$ is nonsingular,

$$\max_{\mathbf{A}^T\mathbf{A}=I_s} \left[\text{trace}((\mathbf{A}^T\mathbf{W}\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{B}\mathbf{A})) - \alpha \sum_{l=1}^{p}\sum_{i=1}^{s} |A_{li}| \right], \quad \alpha \geq 0 \tag{3.29}$$

where α is the degree of sparsity. This problem targets a well-conditioned weight $\mathbf{A}$ with orthogonal columns, using the orthogonal constraint $\mathbf{A}^T\mathbf{A} = I_s$.

Since $\mathbf{W}$ is singular in the case of high-dimensional data, Ng et al. (2011) applied a simple perturbation strategy so that $\mathbf{W}$ is replaced by $\mathbf{W} + \mu\mathbf{I}_p$. Consequently, (3.29) is modified for high-dimensional LDA as

$$\max_{\mathbf{A}^T\mathbf{A}=I_s} \left[\text{trace}((\mathbf{A}^T\mathbf{W}\mathbf{A} + \mu\mathbf{I}_s)^{-1}(\mathbf{A}^T\mathbf{B}\mathbf{A})) - \alpha \sum_{l=1}^{p}\sum_{i=1}^{s} |A_{li}| \right], \quad \alpha \geq 0 \tag{3.30}$$

Finally, to solve problem (3.30), Ng et al. (2011) proposed a gradient flow method with the orthogonal constraint. Suppose a smooth function F is defined on the constraint set $S_t(s,p)$. Then the gradient $\text{grad}(F(\mathbf{A}))$ of F at $\mathbf{A} \in S_t(s,p)$ is given by

$$\text{grad}(F(\mathbf{A})) = \Pi_T\left(\frac{\partial F(\mathbf{A})}{\partial \mathbf{A}}\right) \quad \forall \mathbf{A} \in S_t(s,p) \tag{3.31}$$

where

$$\Pi_T(Z) = \mathbf{A}\left(\frac{\mathbf{A}^TZ - Z^T\mathbf{A}}{2}\right) + (\mathbf{I}_p - \mathbf{A}\mathbf{A}^T)Z \in T_AS_t(s,p) \quad \forall Z \in \mathbb{R}^{p\times s} \tag{3.32}$$

is the orthogonal projection of $Z \in \mathbb{R}^{p\times s}$ onto the tangent space $T_AS_t(s,p)$ at $\mathbf{A}$.

It can be observed that the objective function in (3.30) is not smooth because of the additional term $\alpha\sum_{l=1}^{p}\sum_{i=1}^{s}|A_{li}|$. Consequently, Ng et al. (2011) used the following method to approximate the term globally

$$A_{li} \approx A_{li}(\varepsilon) = \sqrt{A_{li}^2 + \varepsilon^2},$$

where $\varepsilon > 0$ is a very small number. Hence, the derivative of the approximated objective function at $\mathbf{A}$ is given as

$$\frac{\partial F(\mathbf{A})}{\partial \mathbf{A}} = 2[\mathbf{BA} - \mathbf{WA}(\mathbf{A}^T\mathbf{WA} + \mu\mathbf{I}_s)^{-1}(\mathbf{A}^T\mathbf{BA})]\left((\mathbf{A}^T\mathbf{WA} + \mu\mathbf{I}_s)^{-1} - \alpha\left(\frac{\partial \sum_{l=1}^{p}\sum_{i=1}^{s} A_{li}}{\partial \mathbf{A}}\right)\right) \tag{3.33}$$

and the gradient grad$F(\mathbf{A})$ can be easily found by substituting (3.33) into (3.31).

Finally, the gradient flow related to the objective function $F(\mathbf{A})$ is generated by the dynamical systems (Ng et al., 2011) given below

$$\frac{d\mathbf{A}(t)}{dt} = \text{grad}(F(\mathbf{A})) = \Pi_T\left(\frac{\partial F(\mathbf{A})}{\partial \mathbf{A}}\right). \tag{3.34}$$

It is noted that $S_t(s,p) \to \mathbb{R}$ is a critical point of any local maximum (or local minimum) of the function $F(\mathbf{A})$. In addition, the gradient flow $\mathbf{A}(t)$ exists for all $t \geq 0$, and converges to a connected component of the set of critical points of $F(\mathbf{A})$ as $t \to \infty$. They furthermore noted that for any value $\mathbf{A}(0) = \mathbf{A}_0 \in S_t(s,p)$, there is a unique trajectory $\mathbf{A}(t)$ starting from $\mathbf{A}_0$ for $t > 0$.

The importance of sparse LDA using the gradient flow algorithm is because it directly approaches the high-dimensional LDA by assuming objects are found in subspaces. Furthermore, it incorporates a sparsity constraint to identify an important set of variables for classification. However, this method used a perturbation strategy when $\mathbf{W}$ is singular. This strategy is the same as the method of covariance regularization. Hence, it becomes closer to the independence rule as the perturbing parameter μ in (3.30) gets larger. Moreover, this method used an approximation method so as to make the objective function smooth. The effect of this approximation on classification results is not clear.

3.4 Optimal scoring methods

Another equivalent formulation of discriminant analysis is using the regression approach framework. Fisher (1936) has shown that, in binary classification, linear discriminant analysis can be recast as linear regression by treating the p **x**-variables as independent variables and the group indicator vector **y** as a dependent variable. This method was extended to more than two classes by Breiman and Ihaka (1984) for a non-linear discriminant analysis using additive models. This method optimizes the scaling of indicators of classes together with the discriminant functions, and hence it is called optimal scoring (OS) approach (Merchante et al., 2012). The idea of optimal scoring is to recast the discrimination problem as a regression problem in which the categorical variables are turned into quantitative variables by assigning scores to classes (Clemmensen et al., 2011).

Various methods extended the OS approach to high dimensional discriminant analysis. These methods will be briefly reviewed in this section. As preliminary, we redefine some notations as follows. We recall that the multivariate data $\mathbf{X}$ consists of n observations, with each observation $\mathbf{x}_j \in \mathbb{R}^p$ comprises of p-variables. Let $\mathbf{Y}$ denote an $n \times g$ group indicator matrix, with columns that correspond to the dummy-variable codings of the g-groups. That is, $\mathbf{y}_{ij} \in \{0, 1\}$ indicates whether the j^{th} observation belongs to the i^{th} group. We assume that the columns of $\mathbf{X}$ are centered (i.e., orthogonal to the constant vector $\mathbf{1}$) so that the mean will be zero and the total sample covariance matrix will be $\mathbf{S} = n^{-1}\mathbf{X}^T\mathbf{X}$.

3.4.1 Penalized discriminant analysis

Hastie et al. (1995) proposed a penalized version of LDA based on OS in a situation where there are many highly correlated variables. They applied smoothness penalty on the discriminant vectors in the OS problem by incorporating a positive-definite penalty matrix $\boldsymbol{\Omega}$. Hastie et al. (1995) defined the optimal scoring problem in compact form as

$$\begin{aligned} &\min_{\boldsymbol{\theta},\boldsymbol{\beta}} ||\mathbf{Y}\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda(\boldsymbol{\beta}^T\boldsymbol{\Omega}\boldsymbol{\beta}) \\ &\text{subject to } \boldsymbol{\theta}^T\mathbf{Y}^T\mathbf{Y}\boldsymbol{\theta} = \mathbf{I}_s \end{aligned} \tag{3.35}$$

where $\boldsymbol{\theta}$ is a $g \times s$ matrix of scores, and $\boldsymbol{\beta}$ is a $p \times s$ matrix of regression coefficients. The optimal scoring problem (3.35) is equivalent to a penalized LDA when $\mathbf{Y}^T\mathbf{Y}$ and $\mathbf{X}^T\mathbf{X} + \lambda\boldsymbol{\Omega}$ are of full rank. This condition is fulfilled when there are no empty classes and $\boldsymbol{\Omega}$ is positive-definite. To handle situations where this condition is not met, Hastie et al. (1995) replaced the sample within-class covariance matrix ($\mathbf{W}$) by a regularized version $\mathbf{W} + \boldsymbol{\Omega}$; then the LDA proceeds as usual.

The regression coefficient vectors of the OS can be mapped to the corresponding discriminant vectors of the penalized LDA showing the equivalence of OS to the penalized LDA (Hastie et al., 1995). The parameters of this mapping are computed by solving the OS problem (3.35).

The OS optimization problem (3.35) is non-convex. However, it can readily be solved by a decomposition in $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. An algorithm for finding the optimal regression coefficients $\boldsymbol{\beta}^*$ (Hastie et al., 1995; Merchante et al., 2012) has the

following steps:

1. initialize $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^0$ such that $\boldsymbol{\theta}^{0T}\mathbf{Y}^T\mathbf{Y}\boldsymbol{\theta}^0 = \mathbf{I}_s$;

2. compute $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{X}^T\mathbf{Y}\boldsymbol{\theta}^0$;

3. set $\boldsymbol{\theta}^*$ to be the s leading eigenvectors of $\mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{X}^T\mathbf{Y}$;

4. compute the optimal regression coefficients $\boldsymbol{\beta}^* = (\mathbf{X}^T\mathbf{X} + \lambda\boldsymbol{\Omega})^{-1}\mathbf{X}^T\mathbf{Y}\boldsymbol{\theta}^*$.

This approach removes the computational burden of finding eigenvalues, and avoids a costly matrix inversion. It is noted that $(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*)$ are uniquely defined and all critical points are global optima.

The limitation of the OS approach developed by Hastie et al. (1995) is that it does not incorporate sparsity. That is, it does not try to select few discriminant variables among the huge number of variables found in high dimensional discriminant analysis. Hence, the problem of interpretation remains a challenge. Moreover, Hastie et al. (1995) have shown the equivalence of OS to penalized LDA only for binary classification. The connection fails in the general multi-class classification problem.

An alternative OS method using group-lasso was developed by Merchante et al. (2012) shows the equivalence of OS to penalized LDA in a multi-group classification problem. The group-Lasso OS problem is given as

$$\boldsymbol{\beta}_{OS} = \min_{\boldsymbol{\theta},\boldsymbol{\beta}} \frac{1}{2}||\mathbf{Y}\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda \sum_{l=1}^{p} ||\beta_l||_2$$
$$\text{subject to } \boldsymbol{\theta}^T\mathbf{Y}^T\mathbf{Y}\boldsymbol{\theta} = \mathbf{I}_s. \tag{3.36}$$

This is equivalent to the penalized LDA problem

$$\begin{aligned} \boldsymbol{\beta}_{LDA} &= \max_{\boldsymbol{\beta}} \boldsymbol{\beta}^T \mathbf{B} \boldsymbol{\beta} \\ \text{subject to } & \boldsymbol{\beta}^T (\mathbf{W} + \lambda \boldsymbol{\Omega}) \boldsymbol{\beta} = \mathbf{I}_s. \end{aligned} \tag{3.37}$$

Both solutions have the form: $\boldsymbol{\beta}_{LDA} = \boldsymbol{\beta}_{OS}\text{diag}\big((\alpha_k^{-1}(1-\alpha_k^2)^{-1/2})\big)$, where $\alpha \in (0,1)$ is the k^{th} leading eigenvalue of $\mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X}+\lambda\boldsymbol{\Omega})^{-1}\mathbf{X}^T\mathbf{Y}$. However, the theoretical properties of the solution of the group-lasso OS problem are not well known.

3.4.2 Sparse discriminant analysis

Clemmensen et al. (2011) proposed a sparse discriminant analysis based on the optimal scoring interpretation of linear discriminant analysis. They defined the sparse discriminant analysis (SDA) sequentially. The k^{th} SDA solution pair $(\boldsymbol{\theta}_k, \boldsymbol{\beta_k})$ solves the problem

$$\begin{aligned} &\min_{\boldsymbol{\theta}_k, \boldsymbol{\beta}_k} \big(||\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k||^2 + \gamma(\boldsymbol{\beta}_k^T \boldsymbol{\Omega} \boldsymbol{\beta}_k) + \lambda||\boldsymbol{\beta_k}||_1\big) \\ \text{subject to } & \frac{1}{n}\boldsymbol{\theta}_k^T\mathbf{Y}^T\mathbf{Y}\boldsymbol{\theta}_k = 1, \;\; \boldsymbol{\theta}_k^T\mathbf{Y}^T\mathbf{Y}\boldsymbol{\theta}_j = 0 \;\; \text{for all } \; j < k \end{aligned} \tag{3.38}$$

where λ and γ are nonnegative tuning parameters. λ controls the degree of sparsity, i.e., when λ is large, $\boldsymbol{\beta_k}$ becomes more spares.

In general, this method incorporates sparsity which makes interpretation simpler. However, there is no information regarding the effectiveness of the method for classification purposes.

3.4.3 A direct approach to LDA in ultra-high dimensions

Mai et al. (2012) proposed a direct sparse discriminant analysis based on the

least squares formulation of LDA for binary classification. We recall from the classical LDA that when $p < n$, linear discriminant analysis for two-groups classification can be connected to least squares (Fisher, 1936; Mai et al., 2012). However, this connection collapses in high dimensional problems because the sample covariance matrix is singular and the linear discriminant direction is not well defined. As an alternative method for high dimensional problem, Mai et al. (2012) developed a penalized least squares formulation of LDA using the lasso penalty (Tibshirani, 1996).

They used a method for coding the class labels that is the same as the coding method used by Fisher (1936). That is, the two groups are coded as: $y_1 = -n/n_1$ and $y_2 = n/n_2$, where $n = n_1+n_2$. Then the solution to the penalized least squares sparse problem is

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{\beta}} \big(\sum_{i=1}^{n} (y_i - \beta_0 - \mathbf{x}_i\boldsymbol{\beta})^2 + \lambda||\boldsymbol{\beta}||_1 \big). \tag{3.39}$$

Mai et al. (2012) show that the least squares classifier and the LDA rule produce an identical classification. That is, the least squares always estimates the Bayes classification direction, even when the dimension grows faster than any polynomial order of the sample size. This problem can be solved using the least angle regression algorithm (Efron et al., 2004) or the coordinate descent algorithm (Friedman et al., 2007).

The method Mai et al. (2012) focuses on showing the connection between least squares and linear discrimination in high-dimensional problem. This connection exists when least squares is penalized using the lasso. This penalty can also help to achieve sparsity. But, it works only for binary classification.

3.5 Miscellaneous methods

There are techniques that directly estimate discriminant projection directions by minimizing misclassification rate, such as, proposed by Fan et al. (2012) and Cai and Liu (2011). There is also a thresholding method that assumes the population covariance matrix and the mean difference vector are sparse. For example, Shao et al. (2011) considered a sparse LDA by using a thresholding method to estimate the parameters such that the estimated parameter are asymptotically optimal under some conditions. In this section we review these methods.

3.5.1 Regularized optimal affine discriminant (ROAD)

Fan et al. (2012) proposed a method that finds the data projection direction $\mathbf{a}^T = \mu_d^T \mathbf{\Sigma}^{-1}$ by directly minimizing the classification error subject to a capacity constraint on $\mathbf{a}$. They assumed that the correlation between variables has considerable effect on classification and showed that the independence rule performs poor when the variables are positively correlated. They also compared theoretically the Naive Bayes (NB) (i.e., independence rule) and the Fisher discriminant at the population level, and they come to the conclusion that the Fisher discriminant rule performs better than the NB discriminant as ρ deviates away from 0. The objective of their work was to estimate the Fisher discriminant vector $\mathbf{a}$ with reasonable accuracy.

To circumvent the problems in high dimensional discriminant analysis, they proposed a regularized method that selects only the $s(s << p)$ most important variables for classification. In classification, the best s variables are those with the

largest Δ_s, where Δ_s is the counterpart of Δ_p.

By using $\mathbf{a}^T = \mu_d^T \mathbf{\Sigma}^{-1}$, they defined the optimal classifier as: assign $\mathbf{x}$ to π_1 if

$$\delta(\mathbf{x}) = \mathbf{a}^T(\mathbf{x} - \mu) > 0. \tag{3.40}$$

The corresponding associated classification error is

$$W(\delta_F(\mathbf{x})) = 1 - \Phi\Big(\frac{\mathbf{a}^T\mu_d}{(\mathbf{a}^T\mathbf{\Sigma}\mathbf{a})^{1/2}}\Big). \tag{3.41}$$

They assumed that minimizing the misclassification error $W(\delta_F(\mathbf{x}))$ is the same as maximizing $\mathbf{a}^T\mu_d/(\mathbf{a}^T\mathbf{\Sigma}\mathbf{a})^{1/2}$, which is equivalent to minimizing $\mathbf{a}^T\mathbf{\Sigma}\mathbf{a}$ subject to $\mathbf{a}^T\mu_d = \mathbf{1}$. Adding an L_1 constraint for regularization, the problem is written as

$$\mathbf{a}_c = \min_{||\mathbf{a}||_1 \leq c,\, \mathbf{a}^T\mu_\mathbf{d}=1} \mathbf{a}^T\mathbf{\Sigma}\mathbf{a} \tag{3.42}$$

where c controls the degree of sparsity. When it is small, only a few variables will be selected, giving sparsity. There are many ways of regularization in the literature on penalized methods that help to achieve sparsity. The commonly used methods are the Lasso (Tibshirani, 1996), the elastic net (Zou and Hastie, 2005) and other related methods. Fan et al. (2012) called the resulting classifier the regularized optimal affine discriminant (ROAD). They also considered the diagonal ROAD (DROAD) by replacing $D = \text{diag}(\mathbf{\Sigma})$ in (3.42) so as to give comparison with the independence rule.

Using the Lagrangian argument, the problem in (3.42) is reformulated as

$$\bar{\mathbf{a}}_\lambda = \min_{\mathbf{a}^T\mu_\mathbf{d}=1} \Big(\frac{1}{2}\mathbf{a}^T\mathbf{\Sigma}\mathbf{a} + \lambda||\mathbf{a}||_1\Big). \tag{3.43}$$

This optimization problem is a constrained quadratic problem and can be solved by existing methods. However, such methods are slow. Fan et al. (2012) pointed

out that, in the compressed sensing literature, it is common to replace an affine constraint by a quadratic penalty. Based on this idea, problem (3.43) can be approximated as

$$\tilde{\mathbf{a}}_{\lambda,\gamma} = \min\left(\frac{1}{2}\mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} + \lambda||\mathbf{a}||_1 + \frac{1}{2}\gamma(\mathbf{a}^T\mu_{\mathbf{d}} - 1)^2\right). \tag{3.44}$$

For practical purposes, the parameters $\boldsymbol{\Sigma}$ and $\mu_{\mathbf{d}}$ are replaced by their corresponding sample estimates $\hat{\boldsymbol{\Sigma}}$ and $\hat{\mu_{\mathbf{d}}}$, respectively. Fan et al. (2012) also pointed out that $\tilde{\mathbf{a}}_{\lambda,\gamma} \longrightarrow \bar{\mathbf{a}}_{\lambda}$ when $\gamma \longrightarrow \infty$. Moreover, when $\lambda = 0$, the solution $\tilde{\mathbf{a}}_{0,\gamma}$ is always in the direction of $\boldsymbol{\Sigma}^{-1}\mu_{\mathbf{d}}$, the Fisher discriminant direction, regardless of the value of γ.

The minimization problem (3.44) can be solved using a constrained co-ordinate descent algorithm. With this algorithm, the p search directions are just unit vectors $e_1, ..., e_p$, where e_i denotes the i^{th} element in the standard basis of $\mathbb{R}^p$. These unit vectors are used as search directions in each search cycle until some convergence criterion has been met. The procedure for this algorithm is described in detail in Fan et al. (2012).

Finally, Fan et al. (2012) have shown that the sample misclassification error $W(\hat{\delta}(\mathbf{x}))$ is asymptotically equivalent to the oracle misclassification rate $W(\delta(\mathbf{x}))$. They also showed that the Fisher discriminant projection direction converges to the oracle projection direction.

ROAD was developed under the assumption that variables are correlated and it performs well when the variables are really correlated. However, the effectiveness of ROAD is not clear in terms of getting sparser discriminant functions. Moreover, ROAD was developed for two-groups classification and further work

is needed to extend ROAD so that it can be used for classification problems with more than two groups.

3.5.2 A direct estimation approach

In high-dimensional setting, the most commonly used structural assumptions are that $\boldsymbol{\Sigma}$ (or $\boldsymbol{\Sigma}^{-1}$) and the differences of mean vectors μ_d are sparse (Cai and Liu, 2011). Under these assumptions, $\boldsymbol{\Sigma}^{-1}$ and μ_d are estimated separately and are then plugged into Fisher's rule. However, Fisher's discriminant rule depends on the product of $\boldsymbol{\Sigma}^{-1}$ and μ_d, i.e. $\boldsymbol{\Sigma}^{-1}\mu_d$. Cai and Liu (2011) criticize methods that estimate $\boldsymbol{\Sigma}^{-1}$ and $\mu_{\mathbf{d}}$ separately, and argued that the product $\boldsymbol{\Sigma}^{-1}\mu_{\mathbf{d}}$ can be estimated directly and efficiently, even when $\boldsymbol{\Sigma}^{-1}$ and/or $\mu_{\mathbf{d}}$ cannot be well estimated separately.

They estimated this product using an l_1 minimization constraint for sparse LDA as

$$\hat{\mathbf{a}} = \min_{\mathbf{a}} ||\mathbf{a}||_1 \text{ subject to } ||\hat{\boldsymbol{\Sigma}}\mathbf{a} - \hat{\mu_d}||_\infty \leq \lambda \tag{3.45}$$

where $\mathbf{a} := \boldsymbol{\Sigma}^{-1}\mu_d$, and λ is a tuning parameter. The linear programming (3.45) is closely related to the Dantzig selector (Candès and Tao, 2007). They implemented the estimator $\hat{\mathbf{a}}$ using linear programming and named the resulting classifier as "the linear programming discriminant (LPD) rule". LPD rule has computational advantages because it only requires the estimation of a p-dimensional vector via linear programming instead of the estimation of the inverse of a $p \times p$ covariance matrix. The rule performs well when $\boldsymbol{\Sigma}^{-1}\mu_d$ is approximately sparse. This assumption is weaker and more flexible assumption than the assumption that both $\boldsymbol{\Sigma}^{-1}$ and μ_d are sparse Cai and Liu (2011).

The sample misclassification rate of the LPD rule is given as

$$W(\hat{\delta}_{LPD}(\mathbf{x})) = 1 - \frac{1}{2}\Phi\bigg(- \frac{(\hat{\mu} - \mu_{\mathbf{1}})^T\hat{\mathbf{a}}}{(\hat{\mathbf{a}}^T\boldsymbol{\Sigma}\hat{\mathbf{a}})^{1/2}}\bigg) - \frac{1}{2}\Phi\bigg(\frac{(\hat{\mu} - \mu_{\mathbf{2}})^T\hat{\mathbf{a}}}{(\hat{\mathbf{a}}^T\boldsymbol{\Sigma}\hat{\mathbf{a}})^{1/2}}\bigg), \tag{3.46}$$

where ϕ is the normal cumulative distribution function (cdf). Cai and Liu (2011) have shown that the misclassification rate of LPD (3.46) is asymptotically comparable with the oracle misclassification rate under certain conditions. Some of these conditions are that the two samples are of comparable size (i.e. $n_1 \asymp n_2$), the eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$ are bounded from below and above, and Δ_p is bounded away from zero.

Consequently, they specified the regularity conditions as: $n_1 \asymp n_2$, $\log p \leq n$, $c_0^{-1} \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq c_0$ for some constant $c_0 > 0$ and $\Delta_p \geq c_1$ for some $c_1 \geq 0$. Suppose these conditions hold and further let $\lambda = C\sqrt{\Delta_p \log p/n}$ with $C > 0$ a sufficiently large constant, and

$$|\boldsymbol{\Sigma}^{-1}\mu_{\mathbf{d}}|_0 = o\bigg(\sqrt{\frac{n}{\log p}}\bigg).$$

Cai and Liu (2011) showed that $W(\hat{\delta}_{LPD}(\mathbf{x})) - W(\delta_F(\mathbf{x})) \to 0$ in probability as $n \to \infty$ and $p \to \infty$ which shows the consistency of the LPD rule when $\boldsymbol{\Sigma}^{-1}\mu_{\mathbf{d}}$ is sparse. However, in practice the value of the tuning parameter λ is selected by using cross-validation. Moreover, if the above conditions hold and

$$|\boldsymbol{\Sigma}^{-1}\mu_{\mathbf{d}}|_0\Delta_p = o\bigg(\sqrt{\frac{n}{\log p}}\bigg),$$

then

$$\frac{W(\hat{\delta}_{LPD}(\mathbf{x}))}{W(\delta_F(\mathbf{x}))} - 1 = O\bigg(|\boldsymbol{\Sigma}^{-1}\mu_{\mathbf{d}}|_0\Delta_p\sqrt{\frac{\log p}{n}}\bigg) \tag{3.47}$$

with probability greater than $1 - O(p^{-1})$. This shows that a larger Δ_p implies a worse convergence rate for the relative classification error.

Finally, Cai and Liu (2011) noted that when Δ_p is very large, the classification problem is easy and the Bayes misclassification rate can be very small. Thus under this condition, it becomes hard for any data-based classification rule to mimic the performance of the oracle rule.

The direct method proposed by Cai and Liu (2011) is computationally efficient method but the method has good properties only under many assumptions and conditions. In fact, some of the assumptions, such as $\Delta_p \geq c_1$, are quite commonly used in high-dimensional discriminant analysis but the first condition, $n_1 \asymp n_2$, makes this method more limited. Furthermore, the assumption that $\log p \leq n$ is also restrictive. Hence, the method cannot be taken as a general method for discriminant analysis because little is known about its performance when one or more of the conditions are not held.

A similar approach has been proposed by Wang et al. (2013). Their method uses a two-stage LDA for high-dimensional discrimination. It uses l_1 minimization which is linear programming for selecting important variables. This minimization problem has the same formulation as (3.45). Then, the LDA is to be applied on the selected variables. Both methods are similar except the later is a two-stage LDA.

3.5.3 Sparse LDA by thresholding (SLDAT)

Shao et al. (2011) proposed a sparse LDA based on a thresholding methodology for classifying two groups that are normally distributed as $N_p(\mu_{\mathbf{i}}, \mathbf{\Sigma})$ for

$i = 1, 2$, in high dimensional setting. They constructed an LDA that is asymptotically optimal under some sparsity conditions on unknown parameters and some conditions on the divergence rate of p (e.g., $n^{-1} \log p \to 0$ as $n \to \infty$).

The approach uses thresholding estimators of the mean effects $(\mu_{\mathbf{d}})$ and the covariance matrix $(\mathbf{\Sigma})$. Hence, the thresholding procedure to induce sparsity into the estimate of the covariance matrix is given as below

$$\tilde{\Sigma}_{jl} = \hat{s}_{jl} I(|\hat{s}_{jl}| > t_1), \text{ with, } t_1 = M_1 \sqrt{\log p}/\sqrt{n},$$

where M_1 is a positive constant, $\hat{s}_{jl}$ is the $(j, l)^{th}$ element of $\hat{\mathbf{\Sigma}}$, and $I(A)$ is the indicator function of the set A. Letting $M_1 \to \infty$ gives a diagonal estimate of $\mathbf{\Sigma}$, and letting $M_1 = 0$ gives a full estimate of $\mathbf{\Sigma}$. However, Shao et al. (2011) considered the case when only the off-diagonal elements of $\hat{\mathbf{\Sigma}}$ are thresholded. A generalized inverse is used when $\tilde{\mathbf{\Sigma}}$ is not invertible in the thresholding procedure.

Additionally, sparsity is introduced on the mean difference vector by thresholding parameter estimates at a level t_2, where $t_2 = M_2(\log p/n)^{0.3}$, and M_2 is a positive constant. The difference between the means of class i and k is then given as $\tilde{\delta}_{l,ik} = \hat{\delta}_{l,ik} I(|\hat{\delta}_{l,ik}| > t_2)$, where $\hat{\delta}_{ik} = \hat{\mu}_i - \hat{\mu}_k$ and $\hat{\delta}_{l,ik}$ is the l^{th} element of $\hat{\delta}_{ik}$. Therefore, M_1 and M_2 control the degree of diagonalization and the degree of sparsity.

Shao et al. (2011) denoted the misclassification rate of the optimal rule as: $R_{OPT} = \phi(-\Delta_p/2)$, where $0 < R_{OPT} < 1/2$. It can be observed that $R_{OPT} \to 0$ when $\Delta_p \to \infty$ as $p \to \infty$ and $R_{OPT} \to 1/2$ when $\Delta_p \to 0$. The objective of the LDA by thresholding is to find a classification rule (T) such that its associated misclassification rate R_T converges in probability to the same limit as R_{OPT}. It

has been shown that if Δ_p is bounded, then T is asymptotically optimal. Note that the misclassification rate is the same as that given in (3.46).

The sparse LDA by thresholding is a good approach for high dimensional LDA, because it focuses on finding a classification rule by minimizing the misclassification error. Furthermore, it has been shown that the sample misclassification rate associated with the LDA by thresholding is asymptotically the same as the optimal misclassification rate. However, the approach requires several assumptions and conditions. Moreover, they used a shrinkage type of regularization on the covariance matrix and mean difference to achieve sparsity, which neglects the dependence between variables.

There also exist other methods that tackle high dimensional discriminant analysis by minimizing misclassification error. For example, copula discriminant analysis (Han et al., 2013) has been proposed for high dimensional discriminant analysis by incorporating the covariance estimator to classification in a copula model.

3.5.4 Classification using discriminative algorithms

There are other class of classification models such as discriminative models which are used as alternative classification method for high-dimensional data. Discriminative models include machine learning algorithms such as support vector machine (SVM) and kernel regression. The purpose of Machine learning is to represent data as feature vectors and then proceed with training algorithms that seek to optimally partition the feature space into regions.

There are situations where SVM is preferably applied for performing dimen-

sion reduction and classification of high-dimensional data. For instance, Bi et al. (2003) proposed a SVM for dimension reduction using ℓ_1-norm regularization. But the method simply focuses on dimension reduction and gives less attention to the classification accuracy. Haber et al. (2015) proposed a classification by discriminative interpolation framework wherein functional data in the same class are adaptively reconstructed to be more similar to each other. Another discriminative method proposed by Godbole and Sarawagi (2004) classifies text documents into a predefined set of classes.

We can consider the discriminative algorithms as alternative class of methods for dimension reduction in classification. The generative models are typically more flexible than discriminative models in classifying high-dimensional classification problems. Therefore, in this thesis, our focus is to develop generative models, such as sparse discriminant analysis, that effectively deal with discrimination problems when $p >> n$.

3.6 Limitations of the existing high-dimensional discrimination methods

It is important to stress that reducing dimension without taking into account the goal of classification may loose information that could have been useful for discriminating the groups. For instance, while PCA reducing the dimensionality of data, it keeps only the variables associated with the largest eigenvalues. Bouveyron and Brunet-Saumard (2014) explained that the first eigenvectors do not necessarily contain more discriminative information than the other eigenvectors.

Due to the singularity of the within-class covariance matrix in high dimensional discriminant analysis, Fisher's LDA is not directly applicable with high dimensional data. One widely used solution is to assume independence among variables, regardless of the effect of the correlations on classification. This fault is found in sparse discriminant methods that are based on the independence rule, such as the nearest shrunken centroids classifier (Tibshirani et al., 2002), the independence Rule (Bickel and Levina, 2004) and features annealed independence rules (Fan and Fan, 2008). These methods all ignore correlations among variables and thus could lead to irrelevant variable selection and poor classification.

The solution based on regularization may ease computational difficulty, but it gives less attention to variable selection (i.e. sparsity), making results hard to interpret in high-dimensional discriminant analysis. Moreover, all regularization methods require tuning of a parameter and this may not be easy unless cross-validation is used appropriately. Hence, high-dimensional classification requires a method that selects important variables and minimizes classification error simultaneously. Some recent approaches to high-dimensional discriminant analysis are based on the idea of minimizing classification error, as is also types of the methods described in Section 3.5. However, we have seen that almost all such methods work on problems that involve only two groups, and the methods require strong conditions for useful asymptotic to hold.

Having these limitations in mind, there is a need to develop a new sparse LDA as an alternative method to discrimination in high dimensions. In this thesis, we propose some alternative methods of sparse LDA for high dimensional

discriminant analysis. We present the alternative methods in the following chapters.

Chapter 4

Function constrained sparse LDA

4.1 Introduction

In the previous chapter, we reviewed various methods of discriminant analysis for high-dimensional data, noted their various approaches to overcome the singularity problem, and ways in which some of the methods introduce sparseness so that results are interpretable. In this chapter, we assume that the $p \times p$ group covariance matrices are equal. That is, $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \ldots = \mathbf{\Sigma}_g = \mathbf{\Sigma}$. We again let $\mathbf{W}$ be the estimate of the common covariance matrix, $\mathbf{\Sigma}$. For this situation, we propose a new method of discriminant analysis that we call function constrained sparse LDA, which selects very few important variables. In this method, a constrained ℓ_1-minimization penalty is applied to the discrimination problem so as to achieve sparsity. The ℓ_1-minimization is a popular technique to select variables in regression analysis and compressed sensing when $p >> n$. For example, Candès and Tao (2007) used the ℓ_1-minimization penalty with the Dantzig selector to select variables in regression analysis with $p >> n$.

Because of the fact that the within-group covariance matrix is a singular matrix when $p >> n$, $\mathbf{W}^{-1}$ does not exist. To circumvent the singularity problem, we use the diagonal within-group covariance matrix $\mathbf{W}_d = \text{diag}(\mathbf{W})$ in our new sparse LDA method. This is because an estimate of $\mathbf{W}^{-1}$ does not necessarily provide a better classifier. As noted in Section 3.2.1, Fan et al. (2008) showed that LDA cannot be better than random guessing when the number of variables is larger than the sample size due to noise accumulation in estimating the covariance matrix. Their method, Feature Annealed Independence Rules (FAIRs), selects a subset of important features for high-dimensional classification. Another method, developed by Witten and Tibshirani (2011), uses $\mathbf{W}_d$ to select a small number of variables using the Lasso penalty. However, this method fails when p is much larger than n. Hence, the main objective of our method is to find easily interpretable sparse discriminant directions that has better performance in terms of speed and accuracy when compared with other competitive methods in the literature. Because our method selects very few variables, the objective of accuracy sparsity is achieved and the method has good accuracy in examples that we examine.

The chapter is organized as follows. In Section 4.2, some of the motivation for sparse LDA methods are briefly reviewed. In Section 4.3 we propose one new sparse LDA method which we call function constrained sparse LDA (FC-SLDA). We also propose a simplified version of the method called FC-SLDA2 in Section 4.4. The newly proposed methods are illustrated using high-dimensional real data sets and are compared with other exiting methods in Section 4.5. The

results and discussion are presented in Section 4.6. Finally, a short summary of the chapter is given in Section 4.7.

4.2 Sparse Linear Discriminant analysis

Sparse LDA produces linear discriminant functions with only a small number of variables, keeping variables that are useful for discriminating between groups and for identifying group membership of observations. In high-dimensional data analysis, such as most genetic analyses, sparse methods of discrimination ensure better interpretability, greater robustness in the model, and lower computational cost for prediction (Clemmensen et al., 2011; Merchante et al., 2012).

An important procedure in the derivation of sparse LDA is variable selection. In high dimensional data, there are a large number of variables on which measurements have been observed and which are available for analysis. However, only some of these variables may contain information that is useful for the purpose of classification (Rencher, 2002). Consequently, it is necessary to select a set of variables that help discriminate the groups while omitting the other variables that cannot make a significant contribution to the discrimination of the groups, and which may be considered as superfluous/redundent variables. Qiao et al. (2009) pointed out that we do not necessarily increase discriminatory power by increasing the number of variables in the application of Fisher's LDA; instead it leads to overfitting. Some important references for variable selection in high dimensional data are variable selection via the lasso (Tibshirani, 1996), variable selection via the elastic net (Zou and Hastie, 2005), the Dantzig selector (Candès

and Tao, 2007), and the group lasso (Merchante et al., 2012). The traditional approach to sparse LDA is to perform variable selection in a separate step before classification. However, this approach leads to a dramatic loss of information for the purpose of the overall classification problem (Filzmoser et al., 2012). Therefore, there is a need to develop a sparse LDA which performs variable selection and classification simultaneously.

When the number of variables is huge (perhaps tens of thousands), it makes sense to look for methods that produce sparse discriminant functions, i.e. methods that involve only a few of the original variables. Broadly speaking, a vector/matrix is called sparse when it has very few non-zero entries. The number of nonzero entries is called the cardinality of the vector/matrix. There are two main ways to impose sparseness on a vector/matrix solution: by specifying certain cardinality constraints on the solution, or by finding the solution subject to a sparseness inducing penalty. The most popular sparseness inducing penalty is the LASSO (**L**east **A**bsolute **S**hrinkage and **S**election **O**perator), introduced by Tibshirani (1996) for multiple regression problems. For a unit length vector $\mathbf{a}$ ($\|\mathbf{a}\|_2 = 1$), the LASSO has the form $\|\mathbf{a}\|_1 = \sum_i |a_i| < \tau$, where τ is called a tuning parameter. By reducing τ, one forces the smaller entries of $\mathbf{a}$ to become exact zeros. The sparsest $\mathbf{a}$ has only one non-zero entry equal to 1.

It is also possible to obtain a sparse solution by prescribing in advance a pattern of sparseness (Vichi and Saporta, 2009). For example, one can require a sparse matrix $\mathbf{A}$ to have just a single nonzero entry in each row.

Another possible option is to employ vector/matrix majorization (Marshall

and Olkin, 1979). An illustration of the effect of majorization is the following example for unit length vectors from $\mathbb{R}^3_+ = [0,\infty) \times [0,\infty) \times [0,\infty)$:

$$\left(\frac{1}{3},\frac{1}{3},\frac{1}{3}\right) \prec \left(0,\frac{1}{2},\frac{1}{2}\right) \prec (0,0,1)\ ,$$

i.e. the "smallest" vector has equal entries. All of the three components in the first vector are nonzero, the second vector has two nonzero components, and the third vector has only one nonzero component. Note that for any two vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, $\mathbf{v}_1 \prec \mathbf{v}_2$ means that $\mathbf{v}_1$ is Karp reducible to $\mathbf{v}_2$. One can use some procedure for generation of majorization (Marshall and Olkin, 1979, p.128) in order to achieve sparseness. A benefit of such an approach is that sparseness can be achieved without any tuning parameters. For example, the procedure to obtain sparse patterns that was proposed by Trendafilov (1994) is equivalent to what is known now as soft-thresholding. Moreover, the threshold can be found easily by the majorization construction, rather than by tuning different parameters. This form of pattern construction can be further related to the fit, the classification error, and/or other desired features of the solution.

4.3 Function constrained sparse LDA (FC-SLDA)

In modern applications, data often has more variables than observations. Such data are also commonly referred to as small-sample data. Then, the within-scatter matrix is singular and the classical LDA is not defined. Although there exist some proposals to circumvent the problem of high-dimensionality, each of the proposed methods has its own limitation as explained in Section 3.6. Here, we propose an alternative method called function constrained method for sparse

LDA (FC-SLDA).

We use the same notation as in earlier chapters. Thus there are n observations with p variables, and each observation belongs to one of the g groups $(\pi_1, \pi_2, \ldots, \pi_g)$. Also n_i is the number of objects in group π_i, so $\sum_{i=1}^{g} n_i = n$. The between-groups covariance matrix is

$$\mathbf{B} = \frac{1}{g-1} \sum_{i=1}^{g} n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^T$$

and the within-groups covariance matrix is given by

$$\mathbf{W} = \frac{1}{n-g} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (\mathbf{x_{ij}} - \bar{\mathbf{x}}_\mathbf{i})(\mathbf{x_{ij}} - \bar{\mathbf{x}}_\mathbf{i})^T, \tag{4.1}$$

where $\mathbf{x}_{ij}$ is the value of the j^{th}-observation in the i^{th}-group, $\bar{\mathbf{x}}_i$ is the sample mean of the i^{th}-group, and $\bar{\mathbf{x}}$ is the estimate of the overall mean vector $\boldsymbol{\mu}$, $j = 1, \ldots, n_i$, $i = 1, \ldots, g$.

The straightforward idea of replacing the non-existant inverse of $\mathbf{W}$ by some kind of generalized inverse has many drawbacks, and thus is not completely satisfactory. For this reason, Witten and Tibshirani (2011) adopted the idea proposed by Bickel and Levina (2004), which circumvents this difficulty by replacing $\mathbf{W}$ with a diagonal matrix $\mathbf{W}_d$ containing its diagonal, i.e. $\mathbf{W}_d := \mathbf{I}_p \odot \mathbf{W}$, where $\odot$ is an element-wise multiplication. This method penalizes the discriminant vectors in Fisher's discriminant problem and projects the data onto a low dimensional subspace that includes only a subset of the original variables. Note, that Dhillon et al. (2002) were even more extreme and proposed doing LDA for high-dimensional data by simply taking $\mathbf{W} = \mathbf{I}_p$, i.e. PCA of $\mathbf{B}$. Trendafilov and Vines (2009) experimented with this option so as to obtain sparse discriminant functions when $\mathbf{W}$ is singular.

4.3.1 General approach to FC-SLDA

Consider the linear transformation $\mathbf{Y} = \mathbf{A}^\top \mathbf{X}$. The goal of Fisher's LDA is to find a $p \times s$ ($s < p$) transformation matrix $\mathbf{A}$ that produces maximum separation between groups by maximizing the between groups covariance matrix (B) relative to the within-groups covariance matrix ($\mathbf{W}$). The transformation matrix is given as $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_s)$, where $\mathbf{a}_k, k = 1, 2, ..., s$ is the k^{th} column vector of $\mathbf{A}$. We take $\mathbf{A}$ as the matrix that maximizes the Fisher's criterion function $f(\mathbf{A})$:

$$f(\mathbf{A}) = \frac{|\mathbf{A}^T \mathbf{B} \mathbf{A}|}{|\mathbf{A}^T \mathbf{W} \mathbf{A}|} \tag{4.2}$$

where $|.|$ is the determinant.

The maximization problem (4.2) can be rewritten as

$$\max |\mathbf{A}^T \mathbf{B} \mathbf{A}| \quad \text{subject to} \quad |\mathbf{A}^T \mathbf{W} \mathbf{A}| = 1. \tag{4.3}$$

Assume that $\mathbf{A}^T \mathbf{W} \mathbf{A}$ is a diagonalizable matrix such that $\mathbf{A}^T \mathbf{W} \mathbf{A} = \mathbf{I}_s$. Then, $|\mathbf{A}^T \mathbf{W} \mathbf{A}| = 1$ in (4.3).

The matrix $\mathbf{A}$ that maximizes (4.2) is found by solving the generalized eigenvalue problem (4.4)

$$\mathbf{B}\mathbf{A} = \mathbf{W}\mathbf{A}\mathbf{\Lambda}, \tag{4.4}$$

where $\mathbf{\Lambda}$ is the $(s \times s)$ diagonal matrix of the s largest eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ ordered in decreasing order. When $n > p$, the matrix $\mathbf{A}^T \mathbf{W} \mathbf{A}$ is diagonal and it is usual to normalize $\mathbf{A}$ such that $\mathbf{A}^T \mathbf{W} \mathbf{A} = \mathbf{I}_s$.

However, our objective is to find sparse discriminant directions. There are different ways of sparsifying a matrix in high-dimensional cases. In our method, we choose the constrained ℓ_1-minimization penalty to get a sparse matrix $\mathbf{A}$.

So, to find the function constrained sparse LDA, we impose an ℓ_1-minimization on the Fisher's general maximization problem (4.2) as

$$\min ||\mathbf{A}||_1 \quad \text{subject to} \quad \mathbf{A}^T\mathbf{B}\mathbf{A} = \mathbf{\Lambda}_{s\times s}, \;\; \mathbf{A}^T\mathbf{W}\mathbf{A} = \mathbf{I}_s. \tag{4.5}$$

By re-arranging the minimum optimization problem (4.5), the general function constrained sparse LDA (FC-SLDA) problem can be given as:

$$\min_{\mathbf{A}^T\mathbf{W}\mathbf{A}=\mathbf{I}_s} ||\mathbf{A}||_1 + \tau(\mathbf{A}^T\mathbf{B}\mathbf{A} - \mathbf{\Lambda}_{s\times s})^2, \tag{4.6}$$

where $\mathbf{\Lambda}$ is the $(s \times s)$ diagonal matrix of the s largest eigenvalues of $\mathbf{W}_d^{-1}\mathbf{B}$ where $\mathbf{W}_d = \text{diag}(\mathbf{W})$, and τ is a tunning parameter. The ℓ_1 minimization produces sparse linear discriminant functions.

To solve problem 4.6, we have to first transform the problem into an optimization problem with an orthogonal constraint. To find $\mathbf{A}$ that satisfies the orthogonal constraint in problem 4.6, $\mathbf{A}^T\mathbf{W}\mathbf{A}$ has to be transformed using appropriate techniques. Some of the techniques are explained below.

When $n > p$, we can find the square-root matrix of $\mathbf{W}$ using different matrix decomposition methods, such as Cholesky decomposition or QR-decomposition (Seber, 2004). Using the Cholesky decomposition, if we assume that $\mathbf{W}$ is at least a positive semidefinite matrix, it can be decomposed as

$$\mathbf{W} = \mathbf{L}^\top\mathbf{L}$$

where $\mathbf{L}$ is a lower triangular matrix. Then, the constraint $\mathbf{A}^T\mathbf{W}\mathbf{A}$ can be transformed as

$$\mathbf{A}^T\mathbf{W}\mathbf{A} = \mathbf{A}^T\mathbf{L}^\top\mathbf{L}\mathbf{A} = (\mathbf{L}\mathbf{A})^\top(\mathbf{L}\mathbf{A}). \tag{4.7}$$

By letting $\mathbf{U} = \mathbf{LA}$, problem (4.6) can be redefined as

$$\min_{\mathbf{U}^T\mathbf{U}=\mathbf{I}_s} ||\mathbf{U}||_1 + \tau(\mathbf{U}^T\mathbf{L}^{-T}\mathbf{B}\mathbf{L}^{-1}\mathbf{U} - \mathbf{\Lambda}_{s\times s})^2. \tag{4.8}$$

Now it is possible to solve problem (4.7) by using algorithms for constrained ℓ_1-minimization problems with the orthogonal constraint, $\mathbf{U}^T\mathbf{U} = \mathbf{I}_s$. However, our interest in this work is to find a sparse matrix $\mathbf{A}$ in high-dimensions.

Typically the mutual correlations of variables in high-dimensions is of limited important for classification, so we can give less weight to the off-diagonal entries of $\mathbf{W}$. We assume that the interrelationships between variables are less important for classification in high-dimensional small sample size scenarios. As a result, we substitute $\mathbf{W}$ by the diagonal matrix, $\mathbf{W}_d = \text{diag}(\mathbf{W})$.

$\mathbf{W}^{-1}$ does not exist when $p >> n$, but $\mathbf{W}_d$ does have an inverse. Let $\mathbf{U} = \mathbf{W}_d^{1/2}\mathbf{A}$. By applying an approach analogous to the method used with Cholesky decomposition for $n > p$, we use $\mathbf{W}_d^{1/2}$ as the symmetric square-root matrix of $\mathbf{W}_d$ for $p >> n$, because $\mathbf{W}_d^{1/2}\mathbf{W}_d^{1/2} = \mathbf{W}_d$. Now the FC-SLDA problem (4.6) can be reformulated as:

$$\min_{\mathbf{U}^T\mathbf{U}=\mathbf{I}_s} ||\mathbf{U}||_1 + \tau(\mathbf{U}^T\mathbf{W}_d^{-1/2}\mathbf{B}\mathbf{W}_d^{-1/2}\mathbf{U} - \mathbf{\Lambda}_{s\times s})^2. \tag{4.9}$$

This is a PCA-like optimization problem with orthogonal constraint $\mathbf{U}^T\mathbf{U} = \mathbf{I}_s$. Therefore, problem (4.9) can be solved using a method related to the optimization method with orthogonality constraints (Wen and Yin, 2013) or the gradient method (Trendafilov and Jolliffe, 2007) after smoothing the ℓ_1-norm.

But, trying to solve problem (4.9) to find the whole of $\mathbf{A}$ in a single step is not an easy task in high-dimensional discrimination problems as it is computationally expensive. Hence, it is a good idea to find a better way of solving the

problem. One efficient solution sequentially estimates the columns of **A**, one column at a time.

4.3.2 Sequential method of FC-SLDA

We noted in Section 4.3.1 that estimation of the transformation matrix **A** using the general constrained ℓ_1-minimization problem (4.9) is computationally expensive. Hence, we have proposed an efficient method of estimation which sequentially estimates the column vectors of **A** one after the other. These columns of **A** are the discriminant vectors $(\mathbf{a}_1, \ldots, \mathbf{a}_s)$. Let $\mathbf{a}_k$ be the k^{th} discriminant vector. Then replacing **W** by $\mathbf{W}_d$, the maximization problem (2.40) is the same as

$$\max_{\mathbf{a}_k} \frac{\mathbf{a}_k^T \mathbf{B} \mathbf{a}_k}{\mathbf{a}_k^T \mathbf{W}_d \mathbf{a}_k} \quad \text{subject to} \quad \mathbf{a}_k^\top \mathbf{W}_d \mathbf{a}_i = \begin{cases} 1, & i = k, \quad \text{for } i, k = 1, 2, \ldots s \\ 0, & i \neq k. \end{cases} \tag{4.10}$$

To find sparse discriminant vectors, various penalty functions can be imposed on problem (4.10). For example, Trendafilov and Jolliffe (2007) used the Lasso penalty for variable section when $n > p$. We now propose the constrained ℓ_1 penalty for variable selection. Let **a** be the 1^{st} column vector of **A**, then the maximization problem (4.10) is equivalent to

$$\min_{\mathbf{a}} ||\mathbf{a}||_1 \quad \text{subject to} \quad \mathbf{a}^T \mathbf{B} \mathbf{a} = \lambda, \quad \mathbf{a}^T \mathbf{W}_d \mathbf{a} = 1, \tag{4.11}$$

where λ is an eigenvalue of $\mathbf{W}_d^{-1}\mathbf{B}$ that corresponds to the eigenvector **a**.

By rearranging (4.11), the sequential function-constrained reformulation of our sparse LDA is given as:

$$\begin{aligned} \min \quad & \|\mathbf{a}\|_1 + \tau(\mathbf{a}^\top \mathbf{B} \mathbf{a} - \lambda)^2, \\ \mathbf{a}^\top \mathbf{W}_d \mathbf{a} = 1 & \\ \mathbf{a} \perp \mathbf{W}_{i-1} & \end{aligned} \tag{4.12}$$

where $\mathbf{W}_0 = \mathbf{0}_{p\times 1}$ and $\mathbf{W}_{i-1} = \mathbf{W}_d[\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_{i-1}]$, and λ is found as a solution of the standard Fisher's LDA problem with $\mathbf{W} = \mathbf{W}_d$.

Problem (4.12) can be solved using a sequential procedure in which only one discriminant vector is determined at each iteration. However, we can further simplify the minimization problem by putting $\mathbf{b} = \mathbf{W}_d^{1/2}\mathbf{a}$. Then the constraints of the minimization problem (4.12) can be simplified as:

$$\mathbf{a}^T\mathbf{B}\mathbf{a} = \mathbf{b}^\top\mathbf{W}_d^{-1/2}\mathbf{B}\mathbf{W}_d^{-1/2}\mathbf{b},$$

and

$$\mathbf{a}^T\mathbf{W}_d\mathbf{a} = \mathbf{b}^T\mathbf{b}.$$

As $\mathbf{W}_d$ is diagonal, $\mathbf{a}$ and $\mathbf{b}$ have the same sparseness. In general, there is no need to recalculate $\mathbf{a}$ from $\mathbf{b}$, as $\mathbf{a}$ are the raw coefficients and $\mathbf{b}$ are the standardized coefficients of the discriminant functions, which are typically reported in LDA. The standardized coefficients are useful for determining the relative contribution of variables in the separation of groups as explained in Section 4.3.4. Let $\mathbf{b}$ be the i^{th} discriminant vector and λ be the i^{th} eigenvalue of $\mathbf{W}_d^{-1}\mathbf{B}$ associated with $\mathbf{a}_i$, $i = 1, 2, \ldots s$. Then, the modified constrained LDA problem (4.12) for producing sparse discriminant functions is defined as:

$$\begin{aligned} \min \quad & \|\mathbf{b}\|_1 + \tau(\mathbf{b}^\top\mathbf{W}_d^{-1/2}\mathbf{B}\mathbf{W}_d^{-1/2}\mathbf{b} - \lambda_i)^2, \\ \mathbf{b}^\top\mathbf{b} &= 1 \\ \mathbf{b}^\top\mathbf{B}_{i-1} &= \mathbf{0}_{i-1}^\top \end{aligned} \tag{4.13}$$

where τ is a non-negative tuning parameter that controls the sparseness of $\mathbf{b}$, and the matrix $\mathbf{B}_{i-1}$ is composed of all preceding vectors $\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_{i-1}$, that is, $\mathbf{B}_{i-1}$ is the $p \times (i-1)$ matrix defined as $\mathbf{B}_{i-1} = (\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_{i-1})$. The columns in the

solution $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_s)$ are called orthogonal discriminant vectors.

The problem in (4.13) is in fact a function-constraint PCA problem (Trendafilov, 2013). For small data, we can apply a dynamical system approach (Trendafilov and Jolliffe, 2006) to solve (4.13). However, standard algorithms based on the dynamical systems are not directly applicable to (4.13) because the objective function has discontinuous first derivatives due to the inclusion of the ℓ_1-norm. Thus, we should use a smooth approximation of ℓ_1 in the objective function. There are various approximation methods that smooth the ℓ_1-norm.

For example, one method of smoothing the ℓ_1 vector norm is given as:

$$\|\mathbf{b}\|_1 = \mathbf{b}^\top \text{sign}(\mathbf{b}) \approx \mathbf{b}^\top \tanh(\gamma \mathbf{b}) \,, \tag{4.14}$$

with some large $\gamma > 0$.

Another type of smoothing method uses the **epsL** approximation (Wu et al., 2009). It gives:

$$\|\mathbf{b}\|_1 = \sum_{j=1}^{p} |b_j| \approx \sqrt{\mathbf{b}^\top \mathbf{b} + \epsilon}, \tag{4.15}$$

where $\epsilon > 0$ is a very small number. Consequently, the epsL approximation for each of the terms $|b_j|$ is given as $|b_j| \approx \sqrt{b_j^2 + \epsilon}$. Other smoothing options are considered elsewhere (Hage and Kleinsteuber, 2014).

Let f denote the objective function from (4.13), i.e.

$$f(\mathbf{b}) = \|\mathbf{b}\|_1 + \tau(\mathbf{b}^\top \mathbf{W}_d^{-1/2} \mathbf{B} \mathbf{W}_d^{-1/2} \mathbf{b} - \lambda_i)^2. \tag{4.16}$$

This function is differentiable at $\mathbf{b}$ and its solution can be found as an initial value problem for:

$$\frac{d\mathbf{b}_i}{dt} = \mathbf{\Pi}_i \nabla_f(\mathbf{b}_i) \quad , \mathbf{b}_i(0) = \mathbf{b}_i^0 \,, \tag{4.17}$$

where ∇_f denotes the gradient of f with respect to the standard (Frobenius) matrix inner product. That is, the gradient flow related to the objective function $f(\mathbf{b}_i)$ is generated by the dynamical systems (Ng et al., 2011) given below:

$$\frac{d\mathbf{b}_i}{dt} = \text{grad}(f(\mathbf{b})) = \mathbf{\Pi}_i \left(\frac{\partial f(\mathbf{b})}{\partial \mathbf{b}} \right) \tag{4.18}$$

and

$$\mathbf{\Pi}_i = \mathbf{I}_p - \mathbf{B}_i \mathbf{B}_i^T \text{ with } \mathbf{B}_i = [\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_i] \, . \tag{4.19}$$

The current ODE solvers (MATLAB, 2011) are not efficient for solving large optimization problems. They track the whole trajectory defined by the ODE, which is time-consuming and undesirable when only the asymptotic state is of interest (Ng et al., 2011; Trendafilov and Jolliffe, 2007). Therefore, we have developed an algorithm that is appropriate for our minimization problem with orthogonality constraints. Specifically, we have developed an efficient algorithm by improving the gradient method (Trendafilov and Jolliffe, 2007, 2006) and by employing a method for optimization with orthogonality constraints (Wen and Yin, 2013). The main steps of our algorithm are summarized in Section 4.3.3.

The method of (Wen and Yin, 2013) is only appropriate for problems involving the decomposition of full rank matrices. Hence, it cannot be directly applied to our method. A benefit of our algorithm is that it can be applied for any minimization problem with orthogonal constraints.

4.3.3 Algorithm 1: FC-sparse LDA

The steps in the algorithm that implements FC-SLDA are as follows.

1. Let $\mathbf{X}$ be an $n \times p$ grouped multivariate data matrix.

2. Randomly split the data into two sets to form training and test datasets.

3. Find the within-group covariance matrix (**W**) and between group covariance matrix (**B**) of the training dataset defined in (2.39).

4. Form the diagonal matrix $\mathbf{W}_d$ from $\mathbf{W}_d = \text{diag}(\mathbf{W})$.

5. Determine the ordered eigenvalues $\lambda_1, \lambda_2, \ldots \lambda_s$ of $\mathbf{W}_d^{-1}\mathbf{B}$.

6. Set the tuning parameter τ to a positive number , say $0 < \tau \leq 2$.

7. For $k = 1, 2, \ldots s$, find the $p \times 1$ vector $\mathbf{b}_k$ by sequentially solving the problem.

$$\min_{\mathbf{b}_k} \left(\|\mathbf{b}_k\|_1 + \tau(\mathbf{b}_k^\top \mathbf{W}_d^{-1/2} \mathbf{B} \mathbf{W}_d^{-1/2} \mathbf{b}_k - \lambda_k)^2 \right)$$

$$\text{subject to} \quad \mathbf{b}_k^\top \mathbf{b}_i = \begin{cases} 1, & i = k, \quad \text{for } i, k = 1, 2, \ldots s \\ 0, & i \neq k. \end{cases} \tag{4.20}$$

8. Let the solutions of (4.20) in step 7 be $\mathbf{b}_1^*$,$\mathbf{b}_2^*$, $\ldots \mathbf{b}_s^*$.

9. Classify the observations in the training data using $\mathbf{X}\mathbf{b}_1^*, \mathbf{X}\mathbf{b}_2^*, \ldots \mathbf{X}\mathbf{b}_s^*$ and compute the average misclassification error (MCE). Let MCE(τ) is an MCE for a given τ.

10. Change τ and repeat steps 7 and 9 until a value of τ is found that minimizes MCE based on the training data. The final choice of τ's is $\hat{\tau} = \min \text{MCE}(\tau)$. If the minimum is attained at several τ's, the minimum value of these τ's is selected.

11. Denote the final solutions as $\mathbf{b}_1$,$\mathbf{b}_2$, $\ldots \mathbf{b}_s$. Then the discriminant functions are $\mathbf{y}_1 = \mathbf{X}\mathbf{b}_1, \mathbf{y}_2 = \mathbf{X}\mathbf{b}_2, \ldots \mathbf{y}_s = \mathbf{X}\mathbf{b}_s$.

4.3.3.1 Notes on the Algorithm

1. Classification is performed using the usual classification rule of standard LDA. That is, we compute the discriminant scores $\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_s$ and assign each observation to its nearest centroid in this transformed space. Specifically,

   assign the j^{th} observation $\mathbf{x}_j$ to the i^{th} group π_i if

   $$[(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_i)^\top \mathbf{b}_k(\tau)]^2 \leq [(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_l)^\top \mathbf{b}_k(\tau)]^2 \text{ for } i \neq l = 1, 2, \ldots g, j = 1, 2, \ldots n_i. \tag{4.21}$$

   otherwise assign it to another group, where $\hat{\boldsymbol{\mu}}_i$ is the sample mean vector of the i^{th} group, $\hat{\boldsymbol{\mu}}_l$ is the sample mean vector of the l^{th} group, and $\mathbf{b}_k(\tau)$ is the k^{th} discriminant vector which is found by solving problem (4.20) for a given τ.

   Let $I_{ij}^k = 1$ if

   $$[(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_i)^\top \mathbf{b}_k(\tau)]^2 - [(\mathbf{x}_j - \hat{\boldsymbol{\mu}}_l)^\top \mathbf{b}_k(\tau)]^2 \leq 0,$$

   else $I_{ij}^k = 0$. Then the total number of correctly classified observations (n^*) in the training dataset is given as: $n^* = \sum_{i=1}^{g} \sum_{j=1}^{n_i} I_{ij}^k$. Hence the average proportion of misclassified observations, which is equal to the misclassification rate (MCE) for a given τ, is

   $$MCE(\tau) = \frac{n - n^*}{n}. \tag{4.22}$$

   The final choice of τ is $\hat{\tau} = \min MCE(\tau)$. If the minimum is attained at several τ's, the minimum of these τ's is selected.

2. Although τ is a continuous parameter, it is very difficult to consider all values of τ. For simplicity we choose τ in the interval $\tau = 0$ to $\tau = 2$. When $\tau = 0$, the solution vector $\mathbf{b}_k(\tau)$ has only one non-zero entry equal to ± 1. Here the classification is not better than random guessing because the second term in 4.13 switches off, and the solution does not depend on the data. Hence, we choose $\tau > 0$ in the interval $\tau \in [0.1, 0.2, \ldots 1.9, 2.0]$. The algorithm starts at $\tau = 0.1$ and the solution is computed iteratively until we find MCE($\hat{\tau}$) such that MCE($\hat{\tau} - \Delta\tau$) $>$ MCE($\hat{\tau}$) and MCE($\hat{\tau}$) $<$ MCE($\hat{\tau} + \Delta\tau$), where $\Delta\tau = 0.1$. Then we choose τ that has the smallest MCE on the training set.

3. The performance of the resulting discriminant functions is evaluated on the test datasets.

4.3.4 Interpretation and sparseness

Interpretation of a discriminant function is based on the relative importance of the variables in discriminating the groups. Note that, if the original data matrix **X** is not standardized, the coefficients of the linear discriminant function (LDF) are called raw coefficients. The constraint $\mathbf{a}^\top \mathbf{W}_d \mathbf{a}$ in problem (4.12) is diagonal and it is usual to normalize **a** such that $(\mathbf{a}^*)^\top \mathbf{W}_d \mathbf{a}^* = 1$, when the raw coefficients are given as $\mathbf{a}^* = \mathbf{a}(\mathbf{a}^\top \mathbf{W}_d \mathbf{a})^{-1/2}$. This is accomplished by dividing each element of **a** by $(\mathbf{a}^\top \mathbf{W}_d \mathbf{a})^{1/2}$, where **a** is the eigenvector of $\mathbf{W}_d^{-1}\mathbf{B}$.

Let the k^{th} LDF be given by $Y_k = \mathbf{a}_k^{*\top}\mathbf{X}$, where $\mathbf{a}_k^* = (a_{k1}^*, a_{k2}^*, \ldots a_{kp}^*)^\top$, $k = 1, 2, \ldots s$, and $\mathbf{X} = (X_1, X_2, \ldots X_p)^\top$. The contribution of the $X's$ to separation of the groups can be assessed by comparing the raw coefficients a_{kj}^*, $j = 1, \ldots, p$.

However, the use of a discriminant function to assess the relative contribution of the $X's$ to separation of the groups gives meaningful interpretation only if the variables are commeasured, that is, measured on the same scale and with comparable variances. If the variables are not commeasured, we need coefficients $b_{kj}, j = 1, \ldots, p$ that are applicable to standardized variables. Hence, the standardized coefficients must be of the form $b_{kj} = s_j a^*_{kj}, j = 1, \ldots, p$, where s_j is the within-group sample standard deviation of the j^{th} variable obtained as the square-root of the j^{th} element of $\mathbf{W}_d$. In vector form, the standardized coefficients are given as: $\mathbf{b}_k = \mathbf{W}_d^{1/2}\mathbf{a}^*_k, k = 1, 2, \ldots s$.

As $\mathbf{W}_d$ is diagonal, the sparseness of $\mathbf{b}$ depends on the sparseness of $\mathbf{a}^*$. For example, consider a sparse vector $\mathbf{a}^*$ with only two nonzero values out of 10 components. Let $\mathbf{a}^* = (0.5, 0.5, 0, \ldots, 0)^\top$, and $\mathbf{W}_d^{1/2} = \text{diag}(s_1, s_2, s_3, \ldots s_{10})$. Then, the standardized coefficient $(\mathbf{b})$ is calculated as

$$\mathbf{b} = \mathbf{W}_d^{1/2}\mathbf{a}^* = \begin{bmatrix} s_1 & 0 & 0 & \cdots & 0 \\ 0 & s_2 & 0 & \cdots & 0 \\ 0 & 0 & s_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & s_{10} \end{bmatrix} \begin{bmatrix} 0.5 \\ 0.5 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 0.5s_1 \\ 0.5s_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4.23}$$

We can see that $\mathbf{b}$ has only two nonzero components which implies that $\mathbf{a}$ and $\mathbf{b}$ have some equivalence in terms of sparseness. Therefore, we do not need to recalculate $\mathbf{a}^*$ because we use $\mathbf{b}$ for interpretation and the sparseness of $\mathbf{a}^*$ is inherited in $\mathbf{b}$.

4.4 FC-SLDA without eigenvalues (FC-SLDA2)

To further make our method faster, we have also developed a version of sequential FC-SLDA that does not require determination of the eigenvalues of $\mathbf{W}_d^{-1}\mathbf{B}$. We simply call this method "FC-SLDA without λ", and denote it as FC-SLDA2. It can be directly derived from (4.13) as follows.

We know that λ is the maximum value of $(\mathbf{a}^\top\mathbf{B}\mathbf{a})/(\mathbf{a}^\top\mathbf{W}_d\mathbf{a})$ where λ is the largest eigenvalue of $\mathbf{W}_d^{-1}\mathbf{B}$. Hence, any eigenvector of $\mathbf{W}_d^{-1}\mathbf{B}$, say $\mathbf{d} \neq \mathbf{a}$, gives a value smaller than λ. This implies that $\lambda - \lambda_d \geq 0$ where λ_d is the eigenvalue associated with the eigenvector $\mathbf{d}$. So, by substituting λ_d in (4.13) in place of λ, FC-SLDA2 can be formulated as

$$\begin{aligned} \min \quad & \|\mathbf{b}\|_1 + \tau(\mathbf{b}^\top\mathbf{W}_d^{-1/2}\mathbf{B}\mathbf{W}_d^{-1/2}\mathbf{b} - \lambda_d)^2 \,. \\ \mathbf{b}^\top\mathbf{b} &= 1 \\ \mathbf{b}^\top\mathbf{B}_{i-1} &= \mathbf{0}_{i-1}^\top \end{aligned} \tag{4.24}$$

We know that some of the eigenvalues of a singular matrix are zero. So, by letting $\lambda_d = 0$, the simplified form of the second version of function-constrained sparse LDA (FC-SLDA2) is given as:

$$\begin{aligned} \min \quad & \|\mathbf{b}\|_1 + \tau(\mathbf{b}^\top\mathbf{W}_d^{-1/2}\mathbf{B}\mathbf{W}_d^{-1/2}\mathbf{b})^2 \,. \\ \mathbf{b}^\top\mathbf{b} &= 1 \\ \mathbf{b}^\top\mathbf{B}_{i-1} &= \mathbf{0}_{i-1}^\top \end{aligned} \tag{4.25}$$

To solve (4.25), we employ a modified form of the algorithm of FC-SLDA that avoids finding eigenvalues. The advantage of FC-SLDA2 is that it is very fast because it saves the time to calculate the eigenvalues of $\mathbf{W}_d^{-1}\mathbf{B}$. Though it provides less accurate results than FC-SLDA does, FC-SLDA2 is an ideal method for

selecting a small number of variables from an extremely large number of variables. In such a case most of the methods in the literature fail to provide results. For example, the PLDA (Witten et al., 2009) fails to give results when p is very large. Therefore, when we deal with discrimination and classification problems involving, say, tens of thousands of variables or more, the FC-SLDA2 has a practical advantage over most of the commonly used sparse LDA methods available in the literature. The main steps of the algorithm for FC-SLDA2 are given in Algorithm 2 below.

4.4.1 Algorithm 2: FC-SLDA2

The main steps in the algorithm that implements FC-SLDA2 are summarized as follows.

1. Let $\mathbf{X}$ be an $n \times p$ grouped multivariate data matrix.

2. Randomly split the data into two sets to form training and test datasets.

3. Find the within-group covariance matrix ($\mathbf{W}$) and between group covariance matrix ($\mathbf{B}$) of the training data defined in (2.39).

4. Form the diagonal matrix $\mathbf{W}_d$ as $\mathbf{W}_d = \text{diag}(\mathbf{W})$.

5. Set the tuning parameter τ to a positive number , say $0 < \tau \leq 2$.

6. For $k = 1, 2, \ldots s$, find the $p \times 1$ vector $\mathbf{b}_k$ by sequentially solving the prob-

lem.

$$\min_{\mathbf{b}_k} \left(\|\mathbf{b}_k\|_1 + \tau(\mathbf{b}_k^\top \mathbf{W}_d^{-1/2} \mathbf{B} \mathbf{W}_d^{-1/2} \mathbf{b}_k)^2 \right)$$

$$\text{subject to} \quad \mathbf{b}_k^\top \mathbf{b}_i = \begin{cases} 1, & i = k, \quad \text{for } i, k = 1, 2, \ldots s \\ 0, & i \neq k. \end{cases} \tag{4.26}$$

7. Let the solutions of (4.26) in step 7 be $\mathbf{b}_1^*$,$\mathbf{b}_2^*$, $\ldots \mathbf{b}_s^*$.

8. Classify the observations in the training data using $\mathbf{X}\mathbf{b}_1^*, \mathbf{X}\mathbf{b}_2^*, \ldots \mathbf{X}\mathbf{b}_s^*$ and compute the average misclassification error (MCE). Let MCE(τ) be the MCE for a given τ.

9. Change τ and repeat steps 6 and 8 until a value of τ is found that minimizes MCE. The final choice of τ's is $\hat{\tau} = \min \text{MCE}(\tau)$. If the minimum is attained at several τ's, the minimum value of these τ's is selected.

10. Denote the final solutions as $\mathbf{b}_1$,$\mathbf{b}_2$, $\ldots \mathbf{b}_s$. Then the discriminant functions are $\mathbf{y}_1 = \mathbf{X}\mathbf{b}_1, \mathbf{y}_2 = \mathbf{X}\mathbf{b}_2, \ldots \mathbf{y}_s = \mathbf{X}\mathbf{b}_s$.

The tuning parameter (τ) is obtained using the procedures given in Section 4.3.3.1.

In the next section, the newly proposed methods and two other existing prominent methods are each applied to several real datasets and their results compared.

4.5 Numerical applications

We evaluate our method using both small data sets and high-dimensional data sets. We begin with two small data sets in Section 4.5.1 and apply FC-SLDA to high-dimensional data sets in Section 4.5.2.

4.5.1 Applications using small data sets

In this section we evaluate our FC-SLDA methods using two real data sets. The numerical illustrations are given below.

4.5.1.1 Iris data, $n > p$

Iris data (Fisher, 1936) have four variables and three groups with 50 observations in each group. First we applied the original Fisher's LDA (2.40). The effective number of discriminant functions for this problem is $\min(4, 3-1) = 2$. The first two eigenvalues are 32.1919 and 0.2854 (32.4773 in total), and the raw coefficients are depicted in the first two columns of Table 4.1. The projection of the data onto the space spanned by the first two discriminant functions is given in the (1,1) panel of Figure 4.1. It can be seen that there are three misclassified points (52, 103 and 104) for this solution, i.e. 2% misclassification. Then, we solved the original Fisher's LDA with $\mathbf{W} = \mathbf{W}_d$. The first two eigenvalues are 31.0969 and 0.3125 (31.4094 in total), and the raw coefficients are depicted in the second two columns of Table 4.1. There are six misclassified points (9, 31, 50, 52, 103 and 119) for this solution, i.e. 4% misclassification. The discriminant plot of the data is given in the (1,2) panel of Figure 4.1. Next, we solve (4.13) with $\tau = 1.2$. The minimum of the objective function in (4.13) is 1.0680. The first two eigenvalues 31.0969 and 0.3125 are approximated by 30.7763 and 0.4407 respectively. The sparse raw coefficients are given in the third pair of columns in Table 4.1. There are five misclassified points (9, 31, 50, 52, 103) for this solution, i.e. 3.3% misclassification. The discriminant plot of the data is given in the (2,1) panel of Figure 4.1. Finally, we solve (4.13) with $\tau = 0.5$. The minimum of the

objective function in (4.13) is 1.0579. The first two eigenvalues 31.0969 and 0.3125 are approximated by 30.502 and 0.616 respectively. The sparse raw coefficients are depicted in the last pair of columns in Table 4.1. The same five points are misclassified in this solution. The discriminant plot of the data is given in the (2,2) panel of Figure 4.1. It seems that the LDA with $\mathbf{W} = \mathbf{W}_d$ gives the worst solution, while the sparse LDA with $\tau = 0.5$ is most satisfying both in terms of fit and interpretability.

Table 4.1: *Different raw coefficients for Fisher's Iris Data*

Vars	$\mathbf{W}$		$\mathbf{W}_d$		$\text{Sparse}_{1.2}$		$\text{Sparse}_{.5}$	
x_1	-.22	-.31	-.23	-.17	-.17	0	-.15	0
x_2	.28	-.82	.12	-.89	.04	-1.0	0	-1.0
x_3	-.81	.07	-.72	.23	-.74	-.05	-.74	0
x_4	-.46	-.47	-.65	-.35	-.65	0	-.66	0

4.5.1.2 Rice data, $p > n$

Rice data (Krzanowski, 1999; Osborne et al., 1993) have 100 variables (wavelengths) and four groups of rice with 7, 19, 9 and 27 observations in them. The effective number of discriminant functions for this problem is $\min(100, 4-1) = 3$. The first three eigenvalues are 25.3009, 1.6737 and 0.0077, which indicates that the discrimination power of the second and the third discriminant functions are not high. There are 37 misclassified points for this solution, i.e. 37% misclassification. This solution is worse than the results obtained by Krzanowski (1999), who employed PCA as a preprocessing step to reduce the number of variables. The

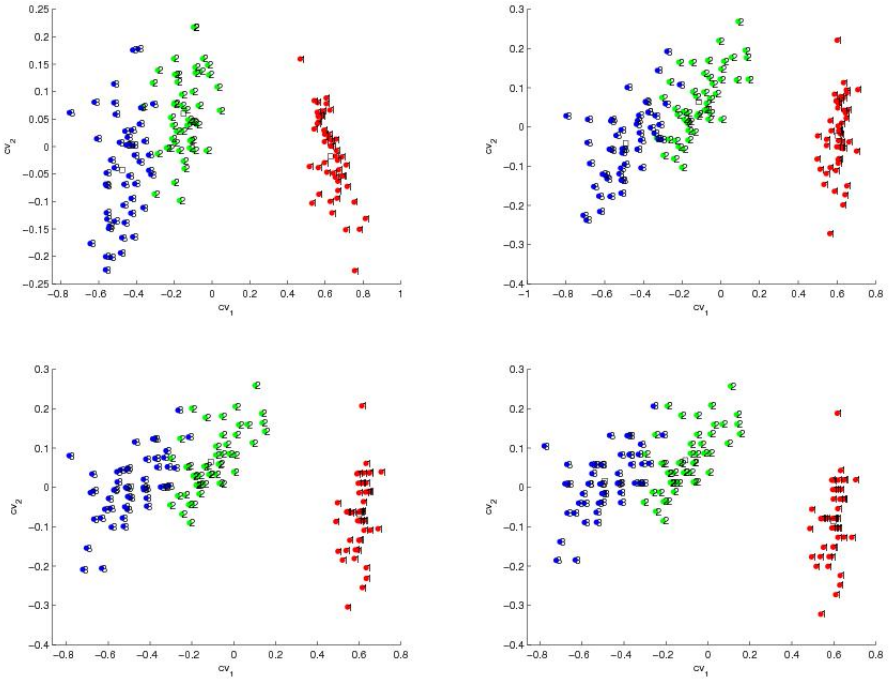

Figure 4.1: *Iris data plotted against two CVs. 1=Iris setosa, 2=Iris versicolor, 3=Iris virginica. Squares denote group means. The* $(1,1)$ *panel uses the original CVs (with* $\mathbf{W}$*). The* $(1,2)$ *panel uses the CVs with* $\mathbf{W}_d$*. The panels* $(2,1)$ *and* $(2,2)$ *use sparse CVs with* $\tau = 1.2$ *and* $\tau = 0.5$ *respectively.*

projection of the data onto the space spanned by the first two discriminant functions is given in the (1,1) panel of Figure 4.2. The panel (1,2) contains the raw coefficients of these discriminant functions. Next, we solve (4.13) with $\tau = 0.5$. The minimum of the objective function in (4.13) is 1.1896. The first three eigenvalues are approximately 23.6843, 0.0874 and 0.0803, respectively. The discriminant plot of the data is given in the (2,1) panel of Figure 4.2. There are 40 misclassified points for this solution, i.e. 40% misclassification. The panel (2,2) contains the raw coefficients of these discriminant functions, and the first ones are not sparse at all. Finally, we solve (4.13) with $\tau = 0.01$. The minimum of the objective func-

tion in (4.13) is 1.0000. The first three eigenvalues are approximately 20.4260, 0.1437 and 0.2418 respectively. The discriminant plot of the data is given in the (3,1) panel of Figure 4.2. There are again 37 misclassified points for this solution, i.e. 37% misclassification. The panel (3,2) contains the sparse raw coefficients of these discriminant functions. It is really surprising to achieve such discrimination using only two variables! The solution is probably too sparse and one might look for a better τ.

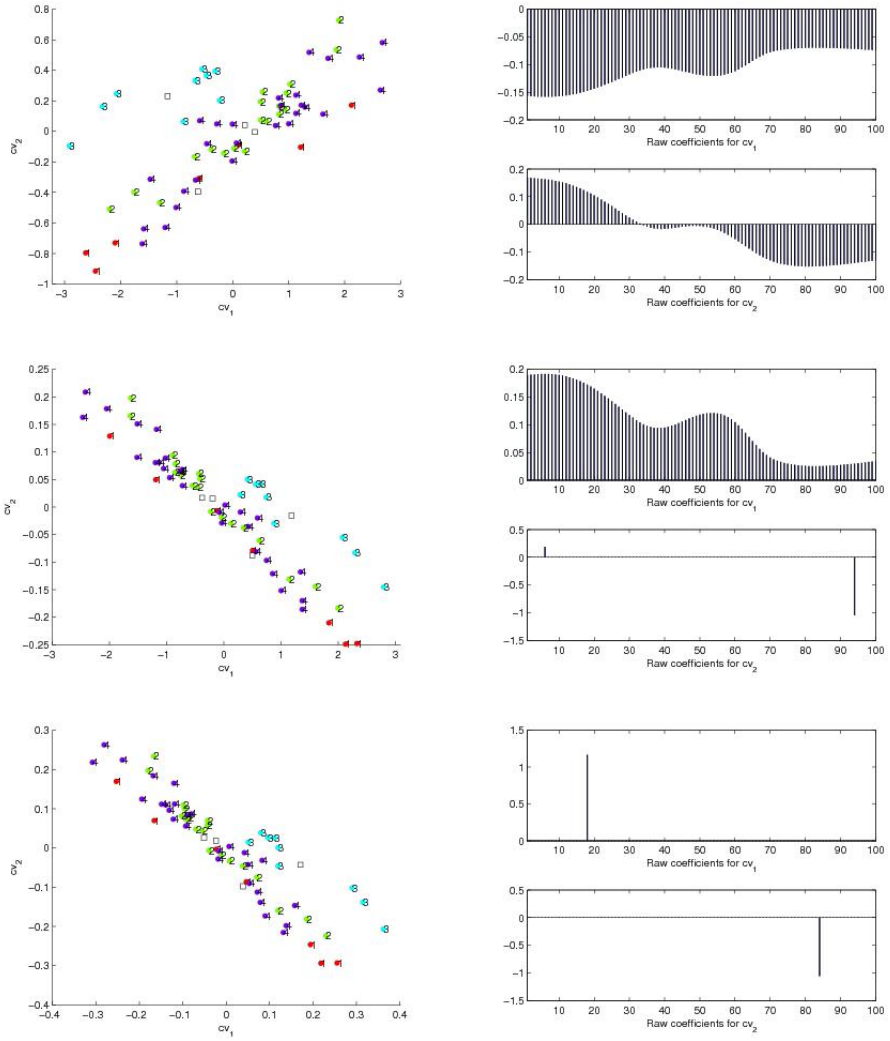

Figure 4.2: *Rice data plotted against two CVs. The groups are 1=France, 2=Italy, 3=India, 4=USA. Squares denote group means. The* $(1,1)$ *panel uses the CVs with* $\boldsymbol{W}_d$*. The panels* $(2,1)$ *and* $(2,2)$*, and* $(3,1)$ *and* $(3,2)$ *use sparse CVs with* $\tau = .5$ *and* $\tau = .01$ *respectively.*

4.5.2 Applications with high-dimensional data

In modern applications the data format often has more variables than observations. Four high-dimensional datasets with $p >> n$ were used to further evaluate the performance of our methods. All of the data are high-dimensional

datasets with $p >> n$. These four datasets are described below.

4.5.2.1 Ramaswamy data

Ramaswamy data is a data set consisting of 16,063 gene expression measurements and 198 samples belonging to 14 distinct cancer subtypes (Ramaswamy et al., 2001). The data set has been studied in several references (see for example Witten and Tibshirani (2011); Witten et al. (2009)) and is available at http://www-stat.stanford.edu/hastie/glmnet/glmnetData/); They were split into a training set containing 75% of the samples and a test set containing 25% of the samples.

4.5.2.2 Leukemia microarray data

Leukemia data were used by Clemmensen et al. (2011) and are available at http://sdmc.i2r.a-star.edu.sg/rp/. The study aimed to classify subtypes of pediatric acute lymphoblastic leukemia. The data consist of 12,558 gene expression measurements for 163 training samples and 85 test samples belonging to 6 cancer classes. The data were analyzed in two steps: a feature selection step was followed by a classification step, using a decision tree structure such that one group was separated using a support vector machine at each tree node.

4.5.2.3 IBD dataset

We further demonstrate the application of our method on the IBD data set examined by Mai et al. (2015). This data set contains 22,283 gene expression levels from 127 people. The people are either normal people, people with Crohns disease or people with ulcerative colitis. The data set can be downloaded from Gene Expression Omnibus with accession number GDS1615. The data sets were randomly split with a 2:1 ratio in a balanced manner to form the training set and

the testing set.

4.5.2.4 Ovarian cancer data

The ovarian cancer data (Conrads et al., 2003) were collected from women who had a high risk of ovarian cancer due to a family or personal history of cancer. The objective is to distinguish ovarian cancer from non-cancer observations. The data contain 216 samples; 121 cancer samples and 95 normal samples. The number of recorded variables were 373,401, but only 4000 variables are considered in this study.

The four data sets are summarized in Table 4.2.

Table 4.2: *Summary of four high-dimensional datasets*

Data	p	n	g	Training sample	Testing Sample
Ramaswamy	16063	198	14	148	50
Leukemia	12558	248	6	163	85
IBD	22283	127	3	85	42
Ovarian Cancer	4000	216	2	144	72

The main difficulty with the data sets in Table 4.2 is that the within-groups covariance matrix is singular and the Fisher's LDA 2.40 is not defined. In addition, the number of variables is huge, and hence we need to use the new methods that can handle singular **W** and produce sparse discriminant functions.

4.6 Results and discussion

We conducted an experiment on each of the four data sets. Each experiment involves evaluating the performance of our two newly proposed methods (FC-SLDA and FC-SLDA2) and two other methods exiting in the literature (PLDA and SDA). Each data set was split into training and test samples. The evaluation of the methods was performed by determining their classification errors on the test samples. The computer time to select the same number of nonzero components in each of the discriminant vectors was also recorded. The classification error (in %) and time (in seconds) of the four methods are summarized in Table 4.3. Note that the classification error and time of each method were found by selecting approximately equal number of variables, except the PLDA which does not select the required number of variables.

4.6.1 Comparison with exiting methods

As noted above, we consider four sparse discriminant analysis methods for comparison using the four data sets presented in Table 4.2. The four methods are:

- Function Constrained Sparse Linear Discriminant Analysis (FC-SLDA), which is introduced in Section 4.3.2;
- Function Constrained Sparse Linear Discriminant Analysis without eigenvalues (FC-SLDA2), which is proposed in Section 4.4;
- Sparse Discriminant Analysis (SDA) which is proposed by Clemmensen et al. (2011). It was reviewed in Chapter 3;

- Penalized Classification using Fisher's Linear Discriminant Analysis (PLDA) which was also reviewed in Chapter 3. This method was proposed by Witten and Tibshirani (2011) for penalizing the discriminant vectors in Fisher's discriminant problem.

In Table 4.3 we summarize the results from numerical experiments with the four methods listed above. For completeness, we also include corresponding results for Fisher's iris data and the rice data.

Table 4.3: *Misclassification rate (in %) and time (in seconds) of four sparse LDA methods. The results were found using the testing data sets.*

Data	FC-SLDA2		FC-SLDA		SDA		PLDA	
	Error	Time	Error	Time	Error	Time	Error	Time
Iris	3.80	0.0012	3.30	0.0013	3.0	0.0013	4.00	0.0120
Rice	37.67	0.0050	37.00	0.0070	37.15	0.0070	38.00	0.0760
IBD	34.63	97.5023	33.50	120.65	30.65	112.2230	34.50	131.0600
Leukemia	31.42	18.2745	22.09	35.3201	27.65	19.9700	27.33	35.2000
Ovarian Cancer	21.05	55.0350	19.03	59.1958	19.31	58.3452	20.65	60.1024
Ramaswamy	18.00	109.3400	13.13	115.1903	16.16	116.5012	–	–

Error denotes misclassification rates as percentages, and Time is the running time of each method in seconds.

The solutions produced by FC-SLDA, FC-SLDA2 and SDA have about 5% non-zero entries for all datasets except the iris data in which two, i.e. 50% vari-

ables were selected to achieve the results. PLDA gives a slightly greater number of nonzero components as compared to the other methods. In addition, PLDA (Witten and Tibshirani, 2011) does not give results for the Ramaswamy data. This may be due to the fact that the Ramaswamy data has a large number of groups, i.e. g=14. So it cannot be compared with the other methods using the Ramaswamy data. We can see that, on each dataset, the proposed FC-SLDA and FC-SLDA2 have reasonably competitive performance in terms of classification errors while selecting few variables. Overall, FC-SLDA performs better than the three other methods in terms of misclassification rates. Though FC-SLDA2 was slightly less accurate than the other methods, it was the fastest method. Hence, in the case of FC-SLDA2, there may be a trade off between accuracy and speed.

4.6.2 Choice of tuning parameter (τ)

The tuning parameter,τ, in the FC-SLDA and FC-SLDA2 methods controls the constraint function. We chose our tuning parameter (τ) for each of the real data sets using the procedures given in Section 4.3.3.1. Therefore, we have chosen the tuning parameter, τ, that gives the lowest classification error. For example, the tuning parameter plotted against classification error of the training dataset of the ovarian cancer data is presented in Figure 4.3.

We can see from Figure 4.3 that the misclassification rate decreases steadily when τ increases from 0 to 0.6. The misclassification rate stabilizes and attains its minimum in the interval 0.6 to 0.9 values of τ. Then the misclassification rate increases again for $\tau \geq 1$. Therefore, we set $\tau = 0.7$ for the ovarian cancer data. We employed similar procedures on the other data sets to choose optimal tuning

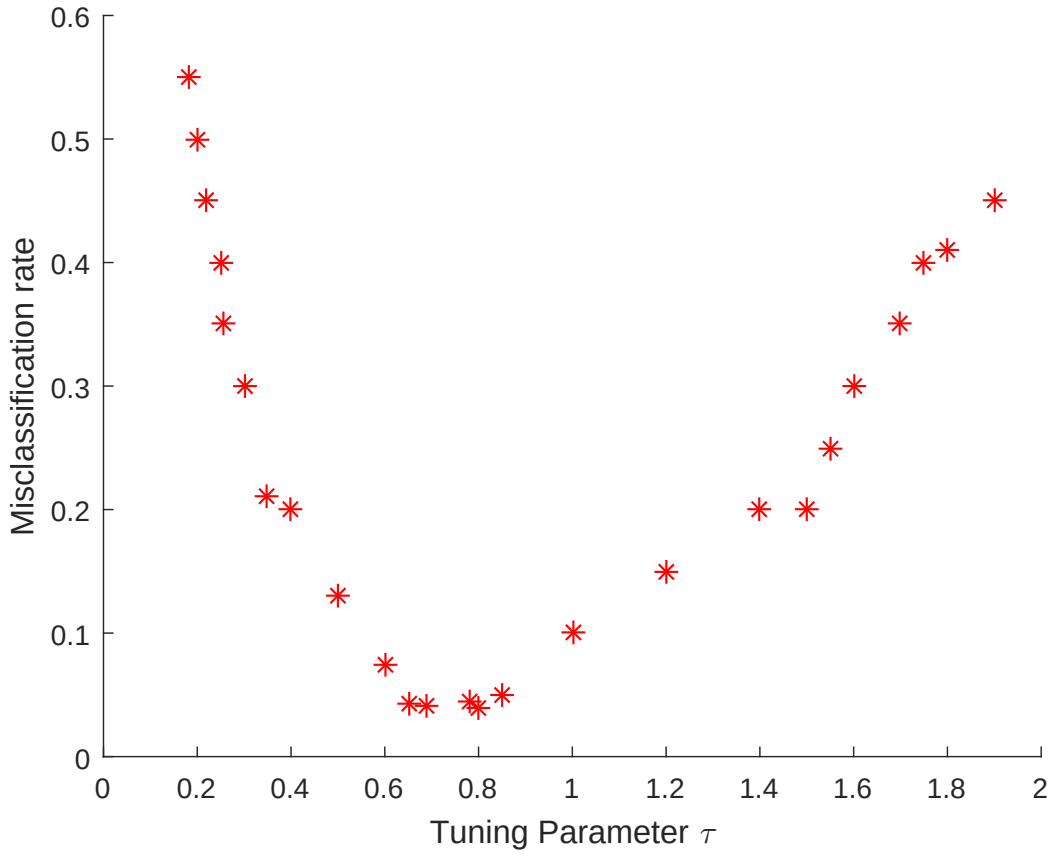

Figure 4.3: *Tuning parameter* (τ) *plotted against misclassification rate for the training data set of the ovarian cancer data. The misclassification rate decreases steadily when* τ *increases from 0 to 0.6. The misclassification rate stabilizes and attains its minimum when* τ *is between 0.6 and 0.9. Then the misclassification rate increases again for* $\tau \geq 1$.

parameters.

4.6.3 Variable selection and sparseness

Our sparse LDA methods select very few non-zero elements gaining good sparseness. We performed the variable selection using cross-validation. The FC-SLDA and FC-SLDA2 select a small number of variables that minimize classification error. Cross validation was performed under the assumption that there is no

interaction between variables. For example, the effective method of sequential variable selection (Fan and Fan, 2008) assumed variables are independent when $p >> n$. Because we have used the diagonal within covariance matrix in developing our method, we employed a similar cross-validation technique used by Fan and Fan (2008).

For illustration, let us again consider the ovarian cancer data. The results of the cross-validation that includes classification error and number of variables used for ovarian cancer data are given in Figure 4.4, which plots the misclassification rate against the number of variables. The cross-validation classification

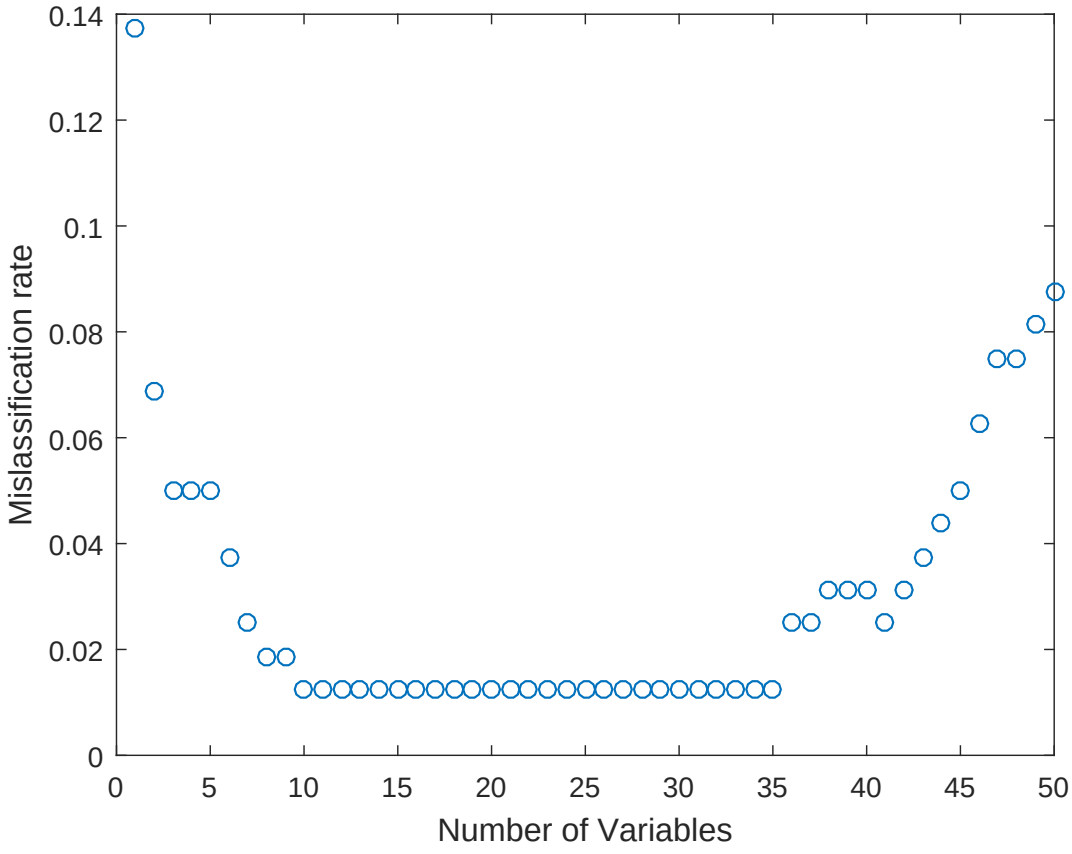

Figure 4.4: *Classification error is plotted against the number of selected variables.*

error reaches minimum when 10 variables are used. The error stays stable over the range from 10 variables to 35 variables. The error goes up when more than 35 variables are used. Therefore, for any number of variables between 10 and 35, the classification error in the non-validation data is minimized. We have also employed the same procedure to select variables for the other data sets. The variable selection technique used achieves the desired spareness. For example, only 10 variables are found useful for efficient classification of the ovarian cancer data. Hence, interpretation is now simpler as we have very few variables.

4.7 Chapter summary

In this chapter a new function constrained sparse LDA (FC-SLDA) and its simplified version (FC-SLDA2) were proposed for high-dimensional discrimination problems. A general method of FC-SLDA was developed to simultaneously find all the column vectors of the discriminant transformation matrix $\mathbf{A}$. However, the general method is computationally expensive. Hence, an efficient sequential method was proposed to iteratively find each discriminant vectors in turn.

An ℓ_1 penalty is employed to find sparse discriminant vectors. This acts as a sparsity penalty in order to select a few variables from a large number of variables. Different high-dimensional real data sets were used to illustrate the methods, and they were compared with two other competitive existing methods. Based on classification error and speed, the results show that FC-SLDA performs well when compared to the other methods. The FC-SLDA2 was found to be the fastest method of discrimination though it has a relatively high classification error.

Chapter 5

Sparse LDA using common principal components

5.1 Introduction

In the previous chapter, we proposed function constrained sparse LDA, and it performs well on high-dimensional real data sets. However, sparse LDA makes the assumption that the different groups share a common within-group covariance matrix. In this chapter we relax these assumptions and allow the within-group covariance matrix to differ between groups but assume some common structure across groups. The first new method proposed in this chapter is called SDCPC-Sparse discriminant analysis with common principal components. This assumes that the principal components do not vary across groups. The other new method proposed in this chapter assumes that the within-groups covariance matrices are proportional to each other. This is equivalent to assuming that they have proportional eigenvalues and common principal components (as well

as sparsity) and we refer to the method as SD-PCPC. The methods are applied to the data sets that were used in Chapter 4 for comparing sparse discriminant methods.

The main assumption in high dimensional discriminant analysis is that the number of variables is too large and, hence, the data at hand actually live in a space of lower dimension, let us say $d < p$. The process of dimension reduction can be done using different variable selection methods (Bouveyron et al., 2007) or PCA (Jolliffe, 2002). A commonly used method is to reduce the dimensionality of the data and then apply classical LDA to the reduced dimension space (Bouveyron et al., 2007; Srivastava and Kubokawa, 2007). That is, once the data are projected into a low-dimensional space, it is possible to apply classical LDA on the projected observations to obtain a partition of the original data. This method is called a two-stage DA. The most common approach is to compute principal components (PCs) of the original variables, and to use them for discrimination. Hotelling (1933) defined PCA as a method that reduces the dimension of the data while keeping as much variation of the data as possible. In other words, PCA aims to find an orthogonal projection of the data set in a low-dimensional linear subspace, such that the variance of the projected data is maximum (Bouveyron and Brunet-Saumard, 2014). This leads to the classical result where the principal axes $(\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_r)$ are the eigenvectors associated with the largest eigenvalues of the sample covariance matrix $\hat{\boldsymbol{\Sigma}}$ of the data.

PCA searches for orthogonal directions $\mathbf{a}$, for which the variance of the projected data $\mathbf{a}^\top \mathbf{x}$ is maximum. Let the sample covariance matrix of $\mathbf{X}$ be $\hat{\boldsymbol{\Sigma}}$, then

the covariance matrix of the projected data $\mathbf{a}^\top \mathbf{x}$ will be $\mathbf{a}^T \hat{\mathbf{\Sigma}} \mathbf{a}$. The criterion for the k^{th} PC direction is given by

$$\max_{\mathbf{a}} \ \mathbf{a}^T \hat{\mathbf{\Sigma}} \mathbf{a} \quad \text{subject to} \ \ \mathbf{a} \perp \mathbf{a}_j, \ \ \text{for} \ j = 1, \ldots, r-1. \tag{5.1}$$

Therefore, the discriminant analysis can be done on the first r score vectors $\hat{\mathbf{a}}_j^\top \mathbf{x}$, $j = 1, \ldots, r$. The number of PCs (r) has to be chosen individually according to a prediction quality criterion, and usually r is much smaller than p (Filzmoser et al., 2012).

An l_1 penalty can be imposed on the objective function (5.1) to find sparse PCA directions. For example, the penalized PCA using the SCoTLASS criterion (Trendafilov and Jolliffe, 2006) is given as:

$$\max \mathbf{a}^T \hat{\mathbf{\Sigma}} \mathbf{a} - \lambda ||\mathbf{a}||_1 \quad \text{subject to} \ \ \mathbf{a} \perp \mathbf{a}_j, \ \ \text{for} \ j = 1, \ldots, r-1, \tag{5.2}$$

where λ controls the degree of sparsity. Now we can obtain score vectors $\mathbf{X}\hat{\mathbf{a}}_k$ for discriminant analysis. However, these methods assume that the within-group covariance matrix is the same for each group. The aim in this chapter is to relax this assumption.

The chapter is organized as follows: it begins by introducing discrimination using common principal components in Section 5.2. The derivation of the general discriminant analysis for CPC is presented in Section 5.3, and sparse LDA using CPC is given in Section 5.4. The numerical illustrations using real data sets are presented in Section 5.5. Finally, sparse LDA using proportional CPC is proposed in Section 5.6.

5.2 Discrimination using common principal components

We aim to develop a technique that allows us to analyze group elements that have common PCs. The estimation of PCs simultaneously in different groups will enable joint dimension reduction. This multi-group PCA is called common principal components (CPC) analysis. Flury et al. (1997) proposed a discrimination method which uses dimension reduction for the purpose of classification by assuming that all differences between two classes occur in a low-dimensional subspace. The additional assumption of CPC is that the spaces spanned by the eigenvectors is identical across the different groups, whereas variances associated with the components are allowed to vary (Flury, 1988). CPC was first introduced to study discriminant problems with different group covariance matrices, but having common principal axes (Flury, 1988; Zou, 2006; Trendafilov, 2010).

Suppose there are g normal groups with mean vector $\boldsymbol{\mu}_i$ and with different covariance matrices $\boldsymbol{\Sigma}_i$, $i = 1, 2, \ldots, g$. The covariance matrix for the i^{th} group can be decomposed as (Flury, 1988; Trendafilov, 2010):

$$\boldsymbol{\Sigma}_i = \mathbf{A}\boldsymbol{\Lambda}_i\mathbf{A}^T, \quad i = 1, \ldots, g, \tag{5.3}$$

where $\boldsymbol{\Sigma}_i$ is a positive definite $p \times p$ population covariance matrix for every i, $\boldsymbol{\Lambda}_i = \text{diag}(\lambda_{i1}, ..., \lambda_{ip})$ is the matrix of eigenvalues and $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_p)$ is an orthogonal $p \times p$ transformation matrix of eigenvectors.

The important assumption of the CPC model is that all covariances matrices $\boldsymbol{\Sigma}_i$'s have the same eigenvectors for each group; the eigenvectors are the columns

of $\mathbf{A}$. We also assume that all λ_i's are distinct. Flury (1988) gives details on how to obtain maximum likelihood estimate of these quantities. The CPC estimation problem (Trendafilov, 2010) is to find the common eigenvectors and corresponding eigenvalues of a given sample covariance matrix $\mathbf{S}_i$, such that equation (5.3) can be redefined as:

$$\mathbf{S}_i \approx \mathbf{A}\mathbf{\Lambda}_i\mathbf{A}^\top, \quad i = 1, \ldots, g, \tag{5.4}$$

where the approximations are as close as possible in some sense.

The common principal axes in g groups ($\mathbf{A}$) and the diagonal matrix $\mathbf{\Lambda}_i = \text{diag}(\mathbf{A}^\top\mathbf{S}_i\mathbf{A})$ can be estimated using maximum likelihood. Flury (1988) has shown that the solutions of the CPC model is given by the generalized system of characteristic equations:

$$\mathbf{a}_j^\top \left(\sum_{i=1}^{g} (n_i - 1) \frac{\lambda_{ij} - \lambda_{im}}{\lambda_{ij}\lambda_{im}} \mathbf{S}_i \right) \mathbf{a}_m = 0, \quad j, m = 1, \ldots, p, \quad j \neq m. \tag{5.5}$$

Problem (5.5) can be solved using

$$\begin{aligned} &\lambda_{ij} = \mathbf{a}_j^\top \mathbf{S}_i \mathbf{a}_j, \quad i = 1, \ldots, g, j = 1, \ldots, p \\ &\text{subject to} \quad \mathbf{a}_j^\top \mathbf{a}_m = \begin{cases} 1, & j = m \\ 0, & j \neq m. \end{cases} \end{aligned} \tag{5.6}$$

Flury (1988) developed an FG-algorithm to estimate $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p)$ and $\mathbf{\Lambda}_i = (\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{ip})$. Many applications of the CPC model, including the estimation of $\mathbf{A}$ for the three group Iris species data, were reported in Flury (1988).

Although the CPC model by Flury (1988) is efficient in estimating $\mathbf{A}$, it fails when $p > n_i$. We know that $\mathbf{S}_i$ is singular when $p > n_i$, and we have $\text{rank}(\mathbf{S}_i) = r < p$.

Let $\mathbf{\Lambda}_i^{(r)}$ be the $p \times p$ diagonal matrix of the first r ranked eigenvalues $\lambda_{i1} \geq \lambda_{i2} \geq \cdots \geq \lambda_{ir} > \lambda_{i,r+1} = \cdots = \lambda_{ip} = 0$. We can write

$$\mathbf{S}_i = \begin{pmatrix} \mathbf{A}_1 & \mathbf{A}_2 \end{pmatrix} \begin{pmatrix} \mathbf{\Lambda}_i^{(r)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{A}_1^\top \\ \mathbf{A}_2^\top \end{pmatrix}, \tag{5.7}$$

where $\mathbf{A}_1$ contains the first r columns of $\mathbf{A}$ corresponding to the non-zero eigenvalues. As a result, we will be using $\mathbf{\Lambda}_i^{(r)}$ and $\mathbf{A}_1$ in place of $\mathbf{\Lambda}_i$ and $\mathbf{A}$, respectively, when $p > n_i$.

When the dimension p is relatively large, information useful for distinguishing the classes is often contained in a few directions $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_r$, where $r < p$. These directions are called the discriminant directions. To find these directions, Zou (2006) proposed a method that is more general than Fisher's linear discriminant analysis but less general than quadratic discriminant analysis. Zou (2006) applied a general likelihood-ratio criterion for measuring the discriminatory power for a given direction $\mathbf{a}$. We will see the derivation of discrimination based on CPC in Section 5.3 below.

5.3 General method for discriminant analysis

We recall from Chapter 2 that Fisher's linear discriminant analysis (LDA) is given as:

$$\max_a \frac{\mathbf{a}^\top \mathbf{B}\mathbf{a}}{\mathbf{a}^\top \mathbf{W}\mathbf{a}} \tag{5.8}$$

where $\mathbf{B}$ is the between-class covariance matrix and $\mathbf{W}$ is the within-class covariance matrix. In fact, given the first $(k-1)$ discriminant directions, the k^{th} direction

is simply given as

$$\max_{a} \frac{\mathbf{a}^\top \mathbf{B}\mathbf{a}}{\mathbf{a}^\top \mathbf{W}\mathbf{a}} \quad \text{subject to} \quad \mathbf{a}^\top \mathbf{W}\mathbf{a}_j = 0 \quad \forall_{j<k}. \tag{5.9}$$

The primary purpose of discriminant analysis is to find linear combinations $\mathbf{a}^T\mathbf{x}$ that have good discriminatory power between classes.

5.3.1 Likelihood approach to discriminant analysis

Fisher's discrimination rule can also be derived using the likelihood method. This alternative way of deriving Fisher's discrimination rule has been proposed by many authors. For example, Zou (2006) considered viewing the discrimination problem from a likelihood framework.

Let us now consider the likelihood approach to develop a general method for discriminant analysis. Suppose $\mathbf{x} \sim f_i(\mathbf{x})$, where $f_i(\mathbf{x})$ is the density function for group i. To examine the separation of groups, hypotheses are defined as:

H_0: The groups are the same

H_1: The groups are not the same.

In this case, the appropriate test statistics for measuring the relative class separation along a fixed direction $\mathbf{a}$ is the (marginal) generalized log-likelihood ratio (LR):

$$LR(\mathbf{a}) = \log \left\{ \frac{\max \prod_{i=1}^{g} \prod_{j=1}^{n_i} f_i^{(\mathbf{a})}(\mathbf{a}^\top \mathbf{x}_{ij})}{\max \prod_{i=1}^{g} \prod_{j=}^{n_i} f^{(\mathbf{a})}(\mathbf{a}^\top \mathbf{x}_{ij})} \right\}, \tag{5.10}$$

where $f_i^{(\mathbf{a})}(.)$ is the marginal density along the projection defined by $\mathbf{a}$ for class i; $f^{(\mathbf{a})}(.)$ is the corresponding density function under the null hypothesis that the classes have the same density function; and $\mathbf{x}_{ij}$ is the j^{th} observation in group i.

As noted in Chapter 2, Fisher's criterion is a special case of $LR(\mathbf{a})$ when $f_i(\mathbf{a})$

is assumed to be normally distributed with mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}$. However, let us first see the derivation of the general discrimination method based on the maximum log-likelihood ratio given in (5.10).

If $f_i(\mathbf{x}) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, the general discriminant method (5.10) can be simplified as below. Under H_0, let $\hat{\boldsymbol{\mu}}$ be the pooled MLE for $\boldsymbol{\mu} = \boldsymbol{\mu}_i$, $i = 1, 2, ., g$, and we know that $\mathbf{S}$, the sample total covariance matrix, is the MLE for $\boldsymbol{\Sigma}$. Under H_1, let $\hat{\boldsymbol{\mu}}_i$ be the MLE for $\boldsymbol{\mu}_i$, and let $\mathbf{S}_i$, be the sample covariance matrix, the MLE for $\boldsymbol{\Sigma}_i$. Then

$$
\begin{aligned}
LR(\mathbf{a}) &= \log \left\{ \frac{\max \prod_{i=1}^{g} \prod_{j=1}^{n_i} \left(\frac{1}{\sqrt{2\pi \mathbf{a}^\top \mathbf{S}_i \mathbf{a}}} \exp \left\{ \frac{-(\mathbf{a}^\top \mathbf{x}_{ij} - \mathbf{a}^\top \hat{\boldsymbol{\mu}}_i)^2}{2\mathbf{a}^\top \mathbf{S}_i \mathbf{a}} \right\} \right)}{\max \prod_{i=1}^{g} \prod_{j=1}^{n_i} \left(\frac{1}{\sqrt{2\pi \mathbf{a}^\top \mathbf{S} \mathbf{a}}} \exp \left\{ \frac{-(\mathbf{a}^\top \mathbf{x}_{ij} - \mathbf{a}^\top \hat{\boldsymbol{\mu}})^2}{2\mathbf{a}^\top \mathbf{S} \mathbf{a}} \right\} \right)} \right\} \\
&= \log \left\{ \frac{(\mathbf{a}^\top \mathbf{S}_1 \mathbf{a})^{-n_1/2} \cdot (\mathbf{a}^\top \mathbf{S}_2 \mathbf{a})^{-n_2/2} \cdot \ldots \cdot (\mathbf{a}^\top \mathbf{S}_g \mathbf{a})^{-n_g/2}}{(\mathbf{a}^\top \mathbf{S} \mathbf{a})^{-n/2}} \times \frac{\exp \left\{ \frac{-\sum_{i=1}^{g} \sum_{j=1}^{n_i} (\mathbf{a}^\top \mathbf{x}_{ij} - \mathbf{a}^\top \hat{\boldsymbol{\mu}}_i)^2}{\mathbf{a}^\top \mathbf{S}_i \mathbf{a}} \right\}}{\exp \left\{ \frac{-\sum_{i=1}^{g} \sum_{j=1}^{n_i} (\mathbf{a}^\top \mathbf{x}_{ij} - \mathbf{a}^\top \hat{\boldsymbol{\mu}})^2}{\mathbf{a}^\top \mathbf{S} \mathbf{a}} \right\}} \right\}
\end{aligned}
\tag{5.11}
$$

$$
\text{Let} \quad f(\mathbf{a}) = \frac{(\mathbf{a}^\top \mathbf{S}_1 \mathbf{a})^{-n_1/2} \cdot (\mathbf{a}^\top \mathbf{S}_2 \mathbf{a})^{-n_2/2} \cdot \ldots \cdot (\mathbf{a}^\top \mathbf{S}_g \mathbf{a})^{-n_g/2}}{(\mathbf{a}^\top \mathbf{S} \mathbf{a})^{-n/2}}. \tag{5.12}
$$

Taking natural logarithm on $f(\mathbf{a})$ gives

$$
\begin{aligned}
\log f(\mathbf{a}) &= \log \left\{ \left((\mathbf{a}^\top \mathbf{S}_1 \mathbf{a})^{-n_1/2} \cdot (\mathbf{a}^\top \mathbf{S}_2 \mathbf{a})^{-n_2/2} \cdot \ldots \cdot (\mathbf{a}^\top \mathbf{S}_g \mathbf{a})^{-n_g/2} \right) \right\} - \log(\mathbf{a}^\top \mathbf{S} \mathbf{a})^{-n/2} \\
&= \frac{n}{2} \log(\mathbf{a}^\top \mathbf{S} \mathbf{a}) - \frac{1}{2} \sum_{i=1}^{g} n_i \log(\mathbf{a}^\top \mathbf{S}_i \mathbf{a}) \\
&= \frac{1}{2} \sum_{i=1}^{g} n_i (\log \mathbf{a}^\top \mathbf{S} \mathbf{a} - \log \mathbf{a}^\top \mathbf{S}_i \mathbf{a}), \quad \text{where} \quad n = \sum_{i=1}^{g} n_i
\end{aligned}
\tag{5.13}
$$

and

$$\begin{aligned} f(C) &= \frac{\exp\left\{\frac{-\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{a}^\top\mathbf{x}_{ij}-\mathbf{a}^\top\hat{\boldsymbol{\mu}}_i)^2}{\mathbf{a}^\top\mathbf{S}_i\mathbf{a}}\right\}}{\exp\left\{\frac{-\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{a}^\top\mathbf{x}_{ij}-\mathbf{a}^\top\hat{\boldsymbol{\mu}})^2}{\mathbf{a}^\top\mathbf{S}\mathbf{a}}\right\}} \\ &= \exp\left\{\sum_{i=1}^{g}\sum_{j=1}^{n_i}\frac{(\mathbf{a}^\top\mathbf{x}_{ij}-\mathbf{a}^\top\hat{\boldsymbol{\mu}})^2}{\mathbf{a}^\top\mathbf{S}\mathbf{a}} - \sum_{i=1}^{g}\sum_{j=1}^{n_i}\frac{(\mathbf{a}^\top\mathbf{x}_{ij}-\mathbf{a}^\top\hat{\boldsymbol{\mu}}_i)^2}{\mathbf{a}^\top\mathbf{S}_i\mathbf{a}}\right\}. \end{aligned} \tag{5.14}$$

Taking natural logarithm on $f(C)$ gives

$$\begin{aligned} \log f(C) &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}\frac{(\mathbf{a}^\top\mathbf{x}_{ij}-\mathbf{a}^\top\hat{\boldsymbol{\mu}})^2}{\mathbf{a}^\top\mathbf{S}\mathbf{a}} - \sum_{i=1}^{g}\sum_{j=1}^{n_i}\frac{(\mathbf{a}^\top\mathbf{x}_{ij}-\mathbf{a}^\top\hat{\boldsymbol{\mu}}_i)^2}{\mathbf{a}^\top\mathbf{S}_i\mathbf{a}} \\ &= \frac{\mathbf{a}^\top\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\hat{\boldsymbol{\mu}})(\mathbf{x}_{ij}-\hat{\boldsymbol{\mu}})^\top\mathbf{a}}{\mathbf{a}^\top\mathbf{S}\mathbf{a}} - \frac{\mathbf{a}^\top\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\hat{\boldsymbol{\mu}}_i)(\mathbf{x}_{ij}-\hat{\boldsymbol{\mu}}_i)^\top\mathbf{a}}{\mathbf{a}^\top\mathbf{S}_i\mathbf{a}} \end{aligned} \tag{5.15}$$

But, the total sample covariance matrix ($\mathbf{S}$) is given as:

$$\mathbf{S} = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\hat{\boldsymbol{\mu}})(\mathbf{x}_{ij}-\hat{\boldsymbol{\mu}})^\top}{n-1} \tag{5.16}$$

and the sample within-group covariance matrix is give as:

$$\mathbf{S}_i = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(\mathbf{x}_{ij}-\hat{\boldsymbol{\mu}}_i)(\mathbf{x}_{ij}-\hat{\boldsymbol{\mu}}_i)^\top}{n-g}, \quad i = 1, 2, \ldots, g. \tag{5.17}$$

Substituting 5.16 and 5.17 into 5.15, $\log f(C)$ is simplified as:

$$\begin{aligned} \log f(C) &= \frac{\mathbf{a}^\top(n-1)\mathbf{S}\mathbf{a}}{\mathbf{a}^\top\mathbf{S}\mathbf{a}} - \frac{\mathbf{a}^\top(n-g)\mathbf{S}_i\mathbf{a}}{\mathbf{a}^\top\mathbf{S}_i\mathbf{a}} \\ &= (n-1) - (n-g) = g-1. \end{aligned} \tag{5.18}$$

We know that

$$\begin{aligned} LR(\mathbf{a}) &= \log(f(\mathbf{a})\cdot f(C)) \\ &= \log f(\mathbf{a}) + \log f(C) \end{aligned} \tag{5.19}$$

Replacing 5.13 and 5.18 into 5.19, we get:

$$LR(\mathbf{a}) = \frac{1}{2}\sum_{i=1}^{g} n_i(\log\mathbf{a}^\top\mathbf{S}\mathbf{a} - \log\mathbf{a}^\top\mathbf{S}_i\mathbf{a}) + g - 1. \tag{5.20}$$

From this, we can see that apart from a constant not depending on **a**

$$LR(\mathbf{a}) \propto \frac{1}{2}\sum_{i=1}^{g} n_i(\log \mathbf{a}^\top \mathbf{S}\mathbf{a} - \log \mathbf{a}^\top \mathbf{S}_i\mathbf{a}). \tag{5.21}$$

We exploit this result to obtain the CPC estimation method for estimating the discriminant vector **a** when data is sparse.

5.4 Sparse LDA based on common principal components

The simplified form of the likelihood-ratio (5.21) is proportional to the following CPC model:

$$\sum_{i=1}^{g} (\frac{n_i}{n})(\log \mathbf{a}^\top \mathbf{S}\mathbf{a} - \log \mathbf{a}^\top \mathbf{S}_i\mathbf{a}), \tag{5.22}$$

where **S** is the total sample covariance matrix.

The objective is to estimate **a** by maximizing (5.22) iteratively. We aim the variability of observations within the same group to be small. Then, groups are more likely to be separated and observations are more likely to be classified correctly. Therefore, we focus on the within-group covariance matrix ($\mathbf{S}_i$) to find the discriminant vector $\mathbf{a}_k$ for $k = 1, 2, ., r$.

Under the CPC model, we recall that $\mathbf{\Lambda}_i = \text{diag}(\mathbf{A}^\top \mathbf{S}_i \mathbf{A})$ where $\mathbf{\Lambda}_i = \text{diag}(\Lambda_{i1}$, $\lambda_{i2}, \ldots, \lambda ir)$ and $\mathbf{A} = \mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r$. Similarly, we can easily show that $\lambda_{ik} = \mathbf{a}_k^\top \mathbf{S}_i \mathbf{a}_k$. Zou (2006) has shown that under the CPC model, if the estimated common eigenvectors $\mathbf{a}_k$ and $\mathbf{a}_j$ are uniformly dissimilar for all $k \neq j$, then the quantity in (5.22) is maximized by the common eigenvector $\mathbf{a}_k$ for which

$$\sum_{i=1}^{g} \left(\frac{n_i}{n}\right)(-\log \lambda_{ik}) \tag{5.23}$$

is the largest.

However, the CPC based discrimination method proposed by Zou (2006) does not show how to estimate each PCs for the purpose of discrimination. Moreover, no similar work exists that incorporates sparsity in such an approach.

We therefore propose a new stepwise estimation method to find the CPCs for discrimination by modifying the CPC estimation method proposed by Trendafilov (2010). This stepwise estimation method imitates standard PCA by finding the CPCs one after another rather than finding all CPCs simultaneously. To find the k^{th} CPC $\mathbf{a}_k$, we solve the following maximization problem:

$$\max_{\mathbf{a}} \sum_{i=1}^{g} \left(\frac{n_i}{n}\right)(-\log \mathbf{a}_k^T \mathbf{S}_i \mathbf{a}_k) \text{ Subject to } ||\mathbf{a}_k||_2^2 = 1 \text{ and } \mathbf{a}^T \mathbf{A}_{k-1} = \mathbf{0}_{k-1}^T. \tag{5.24}$$

This approach is equivalent to Zou (2006)'s approach for maximizing the CPC model in (5.22). Hence, the orthogonal matrix $\mathbf{A} = (\mathbf{a}_1, ..., \mathbf{a}_r)$, which contains the CPCs, is found by solving the maximization problem (5.24) step by step for the k^{th} CPC, $k = 1, 2, ..., r$.

This estimation approach is a very efficient general approach for finding $\mathbf{A}$. However, the method still does not include sparsity. Therefore, we propose to include a lasso-like cardinality constraint on the maximization problem in (5.24) to find sparse results. It is given in Section 5.4.1 below.

5.4.1 Sparsity using a cardinality constraint

By imposing a Lasso penalty (Tibshirani, 1996) on the maximization problem (5.24), we could formulate the sparse LDA using CPC as:

$$\max_{\mathbf{a}} \sum_{i=1}^{g} (\frac{n_i}{n})(-\log \mathbf{a}_k^T \mathbf{S}_i \mathbf{a}_k) - \lambda ||\mathbf{a}_k||_1 \text{ subject to } ||\mathbf{a}_k||_2^2 = 1 \text{ and } \mathbf{a}^T \mathbf{A}_{k-1} = \mathbf{0}_{k-1}^T \tag{5.25}$$

where λ determines the degree of sparsity. The Lasso is more efficient in selecting variables in regression analysis. We assume that the cardinality penalty also performs as efficient as the Lasso. Hence, for simplicity we use the cardinality constraint to select a small number of important variables that are useful for discrimination.

In our method, we impose a cardinality constraint on the maximization problem (5.24), and the resulting sparse LDA using CPC is given as follows.

Let **Card**($\mathbf{a}_k$) be the cardinality (number of non-zero elements) of a vector $\mathbf{a}_k$ and t be an integer with $1 \leq t \leq p$, then the sparse LDA based on CPC is given as:

$$\begin{aligned} &\max_{\mathbf{a}} \sum_{i=1}^{g} (\frac{n_i}{n})(-\log \mathbf{a}_k^T \mathbf{S}_i \mathbf{a}_k) \\ \text{s.t.} \quad &||\mathbf{a}_k||_2^2 = 1, \quad , \mathbf{a}^T \mathbf{A}_{k-1} = \mathbf{0}_{k-1}^T, \quad \mathbf{Card}(\mathbf{a}_k) \leq t. \end{aligned} \tag{5.26}$$

The discriminant vector $\mathbf{a}_k$ is estimated using a stepwise estimation procedure. The first vector to be found is $\mathbf{a}_1$, which gives the maximum of (5.26) on the unit sphere in $\Re^r$. The next vector to be found is $\mathbf{a}_2$, which gives the maximum of (5.26) on the unit sphere $\Re^r$ being orthogonal to $\mathbf{a}_1$. Each vector is found this way until we find a_r.

To select a small number of variables, the cardinality constraint is imposed on the maximization problem to achieve sparsity. Finally, we have developed the SD-CPC algorithm to find sparse discriminant vectors for efficient discrimination. The main steps of the SD-CPC algorithm are given in Section 5.4.2.

5.4.2 Algorithm 3: SDCPC

1. Consider an $n \times p$ grouped multivariate data matrix.

2. Randomly split the data into two sets to form training and testing datasets. Let $\mathbf{X}$ denotes the training data set.

3. For cross-validation, randomly divide the training data into 10 subsets such that each subset contains one tenth of each group.

4. Take nine of the ten subsets and let $\mathbf{X}_{/m}$ denote the data set when the m^{th} subset is omitted and let $\mathbf{X}_c$ denote the omitted data.

5. Put $m = 1$.

6. For the data set $\mathbf{X}_{/m}$, find the covariance matrix for each group ($\mathbf{S}_i$), $i = 1, 2, \ldots g$.

7. Start the cardinality with $t = 1$, where $t < p$.

8. For $k = 1, 2, \ldots, r \leq \min(p, g-1)$, find the $p \times 1$ vector $\mathbf{a}_k$ by solving the problem.

$$\begin{aligned} &\max_{\mathbf{a}} \sum_{i=1}^{g} (\frac{n_i}{n})(-\log \mathbf{a}_k^T \mathbf{S}_i \mathbf{a}_k) \\ &\text{s.t.} \quad \mathbf{a}_k^\top \mathbf{a}_k = 1, \quad \mathbf{a}_k^T \mathbf{A}_{k-1} = \mathbf{0}_{k-1}^T, \quad \mathbf{Card}(\mathbf{a}_k) \leq t. \end{aligned} \tag{5.27}$$

9. Let the solutions of (5.27) be $\mathbf{a}_1^*$,$\mathbf{a}_2^*$, $\ldots$ $\mathbf{a}_r^*$.

10. Classify the observations in the omitted data set, $\mathbf{X}_c$, using the classifiers $\mathbf{X}_c\mathbf{a}_1^*, \mathbf{X}_c\mathbf{a}_2^*, \ldots \mathbf{X}_c\mathbf{a}_r^*$. Record the number of misclassification, calling it Err(m,t).

11. Update t in the interval (1,20] if $p > 20$ and repeat steps 8-10.

12. If $m \leq 10$, increase m by 1 and repeat steps 6-11.

13. Find the value of t that minimizes $\sum_{m=1}^{10} Err(m,t)$. Using all the training data, repeat step 6 for that value of t and let $\mathbf{a}_1$,$\mathbf{a}_2$, $\ldots$, $\mathbf{a}_r$ be the solution to (5.27). The discriminant functions for are $\mathbf{y}_1 = \mathbf{X}\mathbf{a}_1, \mathbf{y}_2 = \mathbf{X}\mathbf{a}_2, \ldots, \mathbf{y}_r = \mathbf{X}\mathbf{a}_r$.

5.4.2.1 Notes on the Algorithm

1. Classification is performed using the usual classification rule of standard LDA. That is, we compute the discriminant scores $\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_r$ and assign each observation to its nearest centroid in this transformed space. Specifically,

   assign an observation $\mathbf{x}$ to the i^{th} group π_i if

$$[(\mathbf{x} - \hat{\boldsymbol{\mu}}_i)^\top \mathbf{a}_k(t)]^2 \leq [(\mathbf{x} - \hat{\boldsymbol{\mu}}_l)^\top \mathbf{a}_k(t)]^2 \text{ for } i \neq l = 1, 2, \ldots g, \tag{5.28}$$

   otherwise assign it to another group, where $\hat{\boldsymbol{\mu}}_i$ is the sample mean vector of the i^{th} group, $\hat{\boldsymbol{\mu}}_l$ is the sample mean vector of the l^{th} group, and $\mathbf{a}_k(t)$ is the k^{th} discriminant vector which is found by solving (5.26) for a given t.

   Let n^* denote the total number of correctly classified observations in the training data set. The proportion of misclassified observations (the misclas-

sification rate) for a given t is

$$\text{MCE}(t) = \frac{n - n^*}{n}. \tag{5.29}$$

2. In order to evaluate the algorithm with real data, the tuning parameter (t) is chosen using the cross-validation from the training data. Then the discriminant vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r$ are determined using just the training data. The discriminant functions are then applied to the test data and the number of misclassifications is recorded and used as a measure for evaluating the algorithm.

5.5 Numerical illustrations

The performance of our new SDCPC algorithm is evaluated based on the 6 real data sets given in Section 4.5, and the results of the analysis are presented in Section 5.5.1. We further compare our method with other existing methods in Section 5.5.2.

5.5.1 Numerical Results of SDCPC on real data sets

We applied our new SDCPC algorithm to the six well known real data sets that were used in Section 4.5. These data sets are:

1. Fisher's Iris data ($n > p$)
2. Rice data ($p > n$)
3. Ovarian Cancer data ($p >> n$)
4. Leukemia data ($p >> n$)

5. Ramaswamy data ($p >> n$)

6. IBD data ($p >> n$)

We analysed the data sets using the new SDCPC algorithm. The summarized numerical results are presented in Table 5.5.1.

Table 5.1: *Numerical results of SDCPC on low and high-dimensional real datasets*

Data	n	p	g	r	t	Error (%)	Time
Iris	150	4	3	2	[2,2]	3.00	0.0019
Rice	62	100	4	3	[3,3,3]	35.48	0.0068
Ovarian Cancer	216	4,000	2	1	[10]	19.33	18.2347
Leukemia	248	12,558	6	3	[15,15,15]	13.17	53.1992
Ramaswamy	198	16,063	14	3	[13,13,13]	32.50	118.4381
IBD	127	22,283	3	2	[13,13]	23.50	105.3508

In the table, **Error** denotes the proportion of misclassified observations in %, **Time** is the average system time in seconds, r is the number of discriminant functions, and t is the number of non-zero components in each vector, $\mathbf{a}_k, k = 1, 2, \ldots, r$. We can see from Table 5.5.1 that the SDCPC performs better with the Iris, Ovarian Cancer, and Leukemia data sets. The Rice and Ramaswamy data sets have relatively higher misclassification rates. This may be due to the fact that the groups in the Rice data are very close to each other, making separation of observations a difficult task (Krzanowski et al., 1995). The relatively weak performance of SDCPC on the Ramaswamy data set may be due to the fact that the Ramaswamy data set has 14 groups, much larger than in the other data

sets. In general, SDCPC was found to be an efficient classification method for high-dimensional multivariate data with $p >> n$ and it selected a small number of variables, which is a plus for interpretation.

Cross-validation (CV) was employed to select the number of variables (i.e., t) using the training data set. With each data set, a 10 fold CV was applied to find t that minimizes misclassification error in the training set. The results for the Leukemia data are presented in Figure 5.1. The figure shows the plot of number of variables against misclassification rates. The misclassification rate (MCE) reaches almost 0.10 when 20 variables are used for classification based on the training data. Therefore, we took the number of variables that minimizes MCE of the training set, which is approximately 20, in Figure 5.1. The MCE attains its minimum when about 15 variables are selected from the test data. Similarly, we employed the same approach to select the number of variables for the other data sets.

To further illustrate the performance of our method a 2-dimensional scatter plot for the IBD data is presented in Figure 5.2. We can see from the plot that the groups are well separated. This suggests that the sparse LDA based on CPC performs efficiently in classification of high-dimensional data.

5.5.2 Comparison with other methods

In this section, we compare our SDCPC method with other exiting methods. The two methods used for comparison are briefly described below.

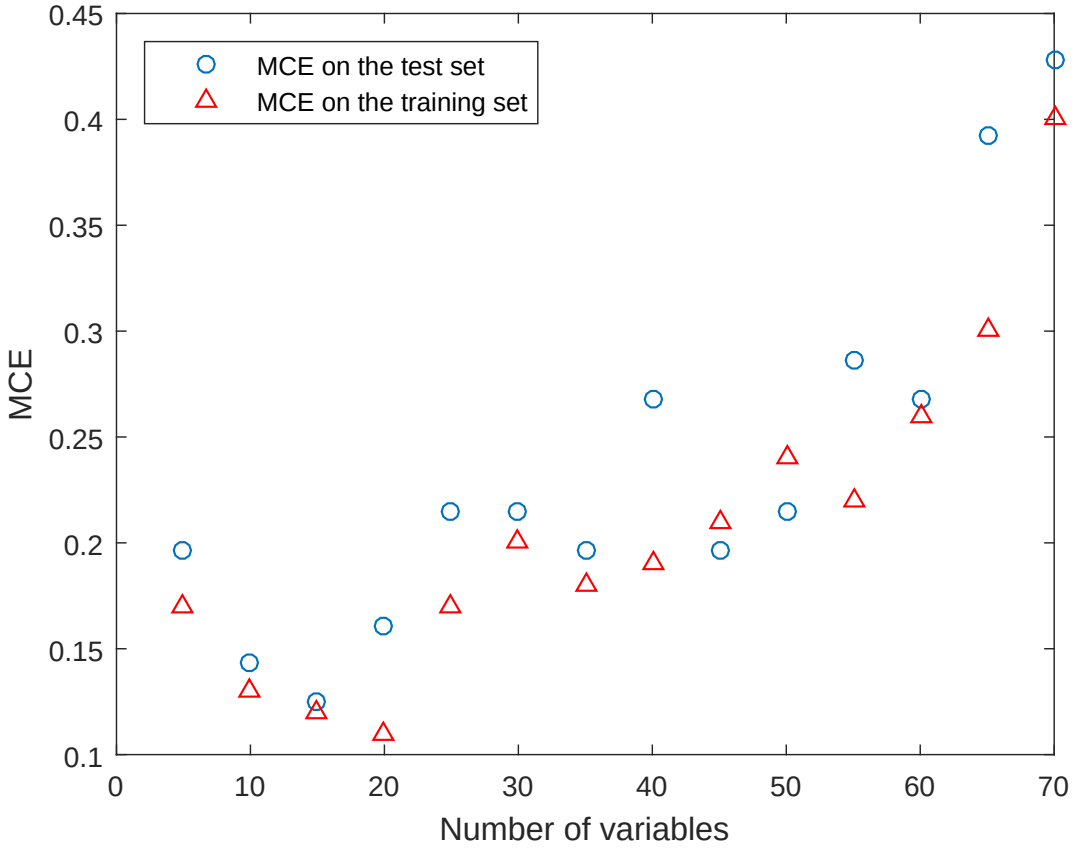

Figure 5.1: *Classification error of training and testing samples is plotted against the number of variables for the Leukemia data.*

5.5.2.1 Penalized linear discriminant analysis (PLDA)

Penalized LDA (Witten and Tibshirani, 2011) penalizes the discriminant vectors in Fisher's discriminant problem. Fisher's discriminant problem finds a low dimensional projection by solving the following problem sequentially

$$\max_{\mathbf{a}_k} \ \mathbf{a}_k^T \mathbf{B} \mathbf{a}_k \text{ subject to } \mathbf{a}^{\mathbf{T}}{}_k \mathbf{W} \mathbf{a}_k \leq 1, \ \mathbf{a}_k^T \mathbf{W} \mathbf{a}_i = 0, \ \forall_i < k.$$

The solution $\mathbf{a}_k$ is the k^{th} discriminant vector $(k = 1, 2, ..., g-1)$. The diagonal estimate of the within-class covariance matrix is used to solve the problem.

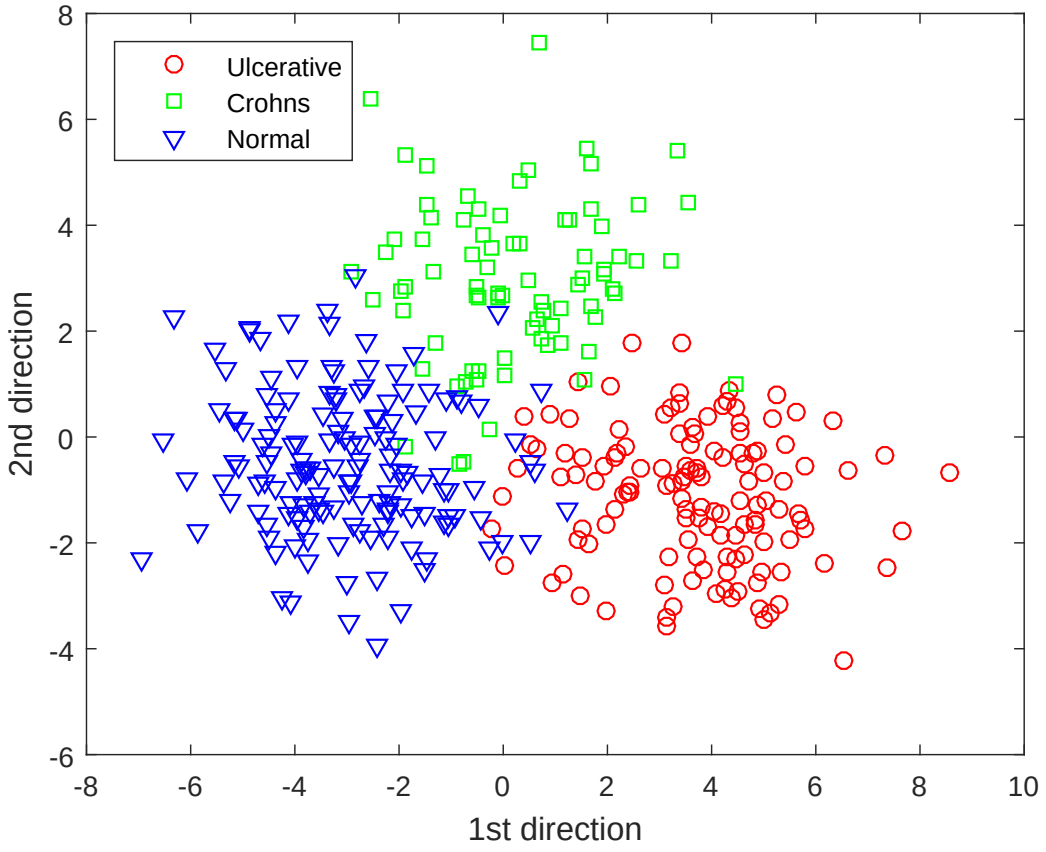

Figure 5.2: *Scatter plot of the three groups of IBD data (i.e. Normal, Crohns, and Ulcerative) using two discriminant directions*

5.5.2.2 Sparse Discriminant Analysis(SDA)

Clemmensen et al. (2011) proposed SDA based on the optimal scoring interpretation of LDA. They defined the sparse discriminant analysis (SDA) method sequentially. Let $\mathbf{Y}$ denote an $n \times g$ group indicator matrix. The k^{th} SDA solution pair $(\boldsymbol{\theta}_k, \boldsymbol{\beta}_k)$ solves the problem

$$\min_{\boldsymbol{\theta}_k, \boldsymbol{\beta}_k} \{||\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k||^2 + \gamma(\boldsymbol{\beta}^T\boldsymbol{\Omega}\boldsymbol{\beta}) + \lambda||\boldsymbol{\beta_k}||_1\}$$

$$\text{subject to } \frac{1}{n}\boldsymbol{\theta}_k^T\mathbf{Y}^T\mathbf{Y}\boldsymbol{\theta}_k = 1, \quad \boldsymbol{\theta}_k^T\mathbf{Y}^T\mathbf{Y}\boldsymbol{\theta}_j = 0 \text{ for } j < k,$$

where λ and γ are nonnegative tuning parameters, and $\mathbf{\Omega}$ is a positive-definite penalty matrix.

5.5.2.3 Results of the three methods based on real data

We compare our method (SDCPC) with the two methodes that are briefly reviewed above, SDA and PLDA. For simplicity, we have taken only two data sets which are randomly selected from the Ovarian Cancer data and IBD data sets for comparison purpose. The two modified data sets are:

1. OC2 data ($n = 216$, $p = 400$, $g = 2$): We took only 400 variables from the total 4000 variables of the Ovarian cancer data.

2. IBD2 data set ($n = 127$, $p = 5000$, $g = 3$): We took only 5000 variables from the 12,283 variables of the IBD data set.

Table 5.2: *Classification error, time and sparsity of three methods*

Data	Criteria	SDCPC	SDA	PLDA
OC2	Errors	18.03	17.31	20.65
	Sparsity (%)	5.0	5.0	5.0
	Time	7.1958	7.0452	10.1024
IBD2	Errors	21.32	21.50	23.50
	Sparsity (%)	2.0	2.0	2.0
	Time	45.1903	40.5012	48.6912

Results of the comparison of the three methods are presented in Table 5.2. Errors denote misclassification rates in percentages, sparsity represents the proportion of non-zero components to the total components, and Time is the running

time of each method in seconds. Based on the modified data sets and their results in Table 5.2, the SDCPC performs better than PLDA in terms of classification and speed with the same sparsity. Our method also provides comparable results with SDA.

Therefore, sparse LDA based on CPC performs effectively in both scenarios (i.e., when $n > p$ and when $p >> n$). The method works with good speed for any size of p. Therefore, sparse LDA based on CPC performs well in classifying observations into their respective groups. Moreover, it gives only a few nonzero components, which helps in identifying the important variables for discrimination.

5.6 Sparse LDA using proportional CPC

The main assumption of classical linear discriminant analysis is that all covariance matrices $\mathbf{\Sigma}_i$(for $i = 1, 2, ..., g$) are identical. However, when the $\mathbf{\Sigma}_i$ are different, quadratic discrimination is an appropriate method. We have also developed two other methods of discrimination based on the structure of the group covariance matrices. These methods are CPC discrimination, which was introduced in Section 5.4 and proportional discrimination. In this section we introduce the discrimination based on proportional CPC. This method is based on the assumption that all $\mathbf{\Sigma}_i$ are proportional (with unknown proportional factors). Replacing the $\mathbf{\Sigma}_i$ in the discrimination rule by their maximum likelihood (ML) estimates or least squares (LS) estimates under proportionality, we find proportional discrimination.

Flury (1988) demonstrated in a simulation study that even a simpler model than CPC with proportional covariance matrices can provide quite competitive discrimination compared to other more complicated methods. For short, we call such PCs proportional PCs (PPC). They are also interesting because they admit very simple and fast implementation that is suitable for large data sets.

As before, we consider g normal populations with mean vector $\boldsymbol{\mu_i}$ and assume that the $p \times p$ covariance matrices, $\boldsymbol{\Sigma}_i$, may be different but are proportional. The hypothesis of proportionality of covariance matrices is given as

$$H_{Prop} : \boldsymbol{\Sigma}_i = c_i \boldsymbol{\Sigma}_1, \quad i = 2, ..., g, \tag{5.30}$$

where c_i are unknown positive constants specific to each population.

We know that under the CPC model, the eigenvalue decomposition (EVD) of $\boldsymbol{\Sigma}_i$ is

$$\boldsymbol{\Sigma}_i = \mathbf{A}\boldsymbol{\Lambda}_i\mathbf{A}^\top, \quad i = 2, ..., g, \tag{5.31}$$

where $\boldsymbol{\Lambda}_i = \text{diag}(\lambda_{i1}, \lambda_{i2}, \ldots, \lambda_{ip})$, and $\mathbf{A}$ is the matrix of common eigenvectors corresponding with $\boldsymbol{\Lambda}_i$. Similarly, let the EVD of $\boldsymbol{\Sigma}_1$ be

$$\boldsymbol{\Sigma}_1 = \mathbf{A}\boldsymbol{\Lambda}_1\mathbf{A}^\top. \tag{5.32}$$

By substituting (5.31) and (5.32) into (5.30), it follows that

$$\boldsymbol{\Lambda}_i = c_i\boldsymbol{\Lambda}_1.$$

As a result, the proportional model can be viewed as an offspring of the CPC Model (Flury, 1988), obtained by imposing the constraints

$$\lambda_{ij} = c_i\lambda_{1j}, \quad i = 1, \ldots, g, \quad j = 1, \ldots, p. \tag{5.33}$$

For simplicity we omit the first index of the diagonal elements of $\mathbf{\Lambda}_1$, that is, we put

$$\mathbf{\Lambda} = \mathbf{\Lambda}_1 = \text{diag}(\lambda_1, \ldots, \lambda_p), \tag{5.34}$$

and the constraints (5.33) are then $\lambda_{ij} = c_i \lambda_j$.

However, when $\mathbf{\Sigma}_1$ is singular, it is replaced by

$$\mathbf{\Sigma}_1 \approx \mathbf{A}\mathbf{\Lambda}_r\mathbf{A}^\top,$$

where $\mathbf{\Lambda}_r = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r)$, and $\mathbf{A}$ is the matrix of common eigenvectors corresponding to the r-nonzero eigenvalues in $\mathbf{\Lambda}_r$. It can be given as $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r)$, where $r < p$.

In the remainder of this chapter we will use the notation $\mathbf{\Lambda}_r$ and $\mathbf{A}$ as the matrices of eigenvalues and their associated eigenvectors, respectively,when dealing with singular covariance matrices.

Therefore, the ML and LS methods are solved under the constraints $\mathbf{A}^\top\mathbf{A} = \mathbf{I}_r$ and $c_1 = 1$. The ML and LS estimation methods of the proportional principal components (PCs) are given in the following sections.

5.6.1 Maximum Likelihood estimation of proportional PCs

Flury (1988) has derived an ML estimation method for proportional PCs. By considering (5.30), the ML estimation of $\mathbf{\Sigma}_i$, i.e. of $\mathbf{\Sigma}_1$ and c_i, is formulated as the following optimization problem:

$$\min_{\Sigma_1, c} \sum_{i=1}^{g} n_i \{\log[\det(c_i\mathbf{\Sigma}_1)] + \text{trace}[(c_i\mathbf{\Sigma}_1)^{-1}\mathbf{S}_i]\}, \tag{5.35}$$

where S_i are given sample covariance matrices and $c = (c_1, c_2, ..., c_g) \in \mathbb{R}^g$ assuming $c_1 = 1$.

Then, after substitution of $\mathbf{\Sigma}_1$, (5.35) becomes:

$$\min \sum_{i=1}^{g} n_i \{\log[\det(c_i \mathbf{A}\mathbf{\Lambda}_r \mathbf{A}^\top)] + \text{trace}[(c_i \mathbf{A}\mathbf{\Lambda}_r \mathbf{A}^\top)^{-1} \mathbf{S}_i]\}, \tag{5.36}$$

which further simplifies to:

$$\min_{A,\lambda,c} \sum_{i=1}^{g} n_i \left\{ \sum_{j=1}^{r} \left[((c_i \lambda_j) + \frac{\mathbf{a}_j^\top \mathbf{S}_i \mathbf{a}_j}{c_i \lambda_j}) \right] \right\}, \tag{5.37}$$

where $\mathbf{a}_j$ and λ_j are respectively the j^{th} eigenvector and eigenvalue of $\mathbf{\Sigma}_1$.

The ML estimates of $\mathbf{a}_j$, λ_j and c_i, are derived from the first order optimality conditions of (5.37). That is, (5.37) can be solved using patrial derivatives with respect to $\mathbf{a}_j$, λ_j and c_i. The detailed procedures of the ML estimation of $\mathbf{a}_j$, λ_j and c_i are given in Flury (1988) for positive definite covariance matrices $\mathbf{\Sigma}_i$, $i = 1, \ldots, p$. They are further used to construct an algorithm for their estimation. However, for high-dimensional multivariate data, the estimation of PPC using the ML algorithm was found to be very slow. Hence, we propose a new least square (LS) estimation method of $\mathbf{a}_j$, λ_j and c_i for high-dimensional discrimination problem. The LS estimation method is presented in Section 5.6.2.

5.6.2 Least square estimation of proportional CPC

We assume that under proportional CPC model, the parameters c_i, $\mathbf{A}$, and $\mathbf{\Lambda}_r$ in (5.32) can be estimated by minimizing the sum of the square of the deviations between $\mathbf{S}_i$ and $c_i \mathbf{A}\mathbf{\Lambda}_r \mathbf{A}^\top$. Therefore, we define the least square (LS) setting of the proportional CPC problem as:

$$\min_{A,\lambda,c} \sum_{i=1}^{g} n_i ||\mathbf{S}_i - c_i \mathbf{A}\mathbf{\Lambda}_r \mathbf{A}^\top||_F^2. \tag{5.38}$$

To find LS estimations of $\mathbf{A}$, $\mathbf{\Lambda}_r$ and $c_2, ..., c_g$ assuming $c_1 = 1$, consider the objective function of (5.38) by letting $Y_i = \mathbf{S}_i - c_i\mathbf{A}\mathbf{\Lambda}_r\mathbf{A}^\top$

$$f = \frac{1}{2}\sum_{i=1}^{g} n_i||Y_i||_F^2 = \frac{1}{2}\sum_{i=1}^{g} n_i \text{trace}(Y_i^\top Y_i), \tag{5.39}$$

and its total derivative:

$$\begin{aligned} df &= \frac{1}{2}d\sum_{i=1}^{g} n_i \text{trace}(Y_i^\top Y_i) = -\sum_{i=1}^{g} n_i \text{trace}[Y_i d(c_i\mathbf{A}\mathbf{\Lambda}_r\mathbf{A}^\top)] \\ &= -\sum_{i=1}^{g} n_i \text{trace}\{Y_i[(dc_i)\mathbf{A}\mathbf{\Lambda}_r\mathbf{A}^\top + c_i\mathbf{A}(d\mathbf{\Lambda}_r) + 2c_i\mathbf{A}\mathbf{\Lambda}_r(d\mathbf{A})^\top]\}. \end{aligned}$$

Then the partial gradients with respect to $\mathbf{A}$, $\mathbf{\Lambda}_r$ and $c_i, i = 2, ..., g$, are:

$$\nabla_{c_i} f = -n_i \text{trace}(Y_i\mathbf{A}\mathbf{\Lambda}_r\mathbf{A}^\top) = n_i c_i \text{trace}(\mathbf{\Lambda}_r^2) - n_i \text{trace}(\mathbf{A}^\top\mathbf{S}_i\mathbf{A}\mathbf{\Lambda}_r) \tag{5.40}$$

$$\nabla_{\Lambda_r} f = -\sum_{i=1}^{g} n_i c_i \mathbf{A}^\top Y_i\mathbf{A} = \sum_{i=1}^{g} n_i c_i^2 \mathbf{\Lambda}_r - \sum_{i=1}^{g} n_i c_i \text{diag}(\mathbf{A}^\top\mathbf{S}_i\mathbf{A}). \tag{5.41}$$

$$\nabla_{\mathbf{A}} f = -2\sum_{i=1}^{g} n_i c_i Y_i\mathbf{A}\mathbf{\Lambda}_r = 2\sum_{i=1}^{g} n_i c_i^2 \mathbf{A}\mathbf{\Lambda}_r^2 - 2\sum_{i=1}^{g} n_i c_i \mathbf{S}_i\mathbf{A}\mathbf{\Lambda}_r. \tag{5.42}$$

At the minimum of (5.39), the partial gradients (5.40) and (5.41) must be zero, which leads to the following LS estimations:

$$c_i = \frac{\text{trace}(\mathbf{A}^\top\mathbf{S}_i\mathbf{A}\mathbf{\Lambda}_r)}{\text{trace}(\mathbf{\Lambda}_r^2)} = \frac{\sum_{j=1}^{r} \mathbf{a}_j^\top\mathbf{S}_i\mathbf{a}_j\lambda_j}{\sum_{j=1}^{r}\lambda_j^2}, \quad i = 2, 3, ...g, \tag{5.43}$$

$$\mathbf{\Lambda}_r = \frac{\sum_{i=1}^{g} n_i c_i \text{diag}(\mathbf{A}^\top\mathbf{S}_i\mathbf{A})}{\sum_{i=1}^{g} n_i c_i^2} \quad \text{or} \quad \lambda_j = \mathbf{a}_j^\top \left(\frac{\sum_{i=1}^{g} n_i c_i \mathbf{S}_i}{\sum_{i=1}^{g} n_i c_i^2}\right) \mathbf{a}_j. \tag{5.44}$$

The gradient (5.42) together with the constraint $\mathbf{A}^\top\mathbf{A} = \mathbf{I}_r$ imply that at the minimum of (5.39) the matrix:

$$\mathbf{A}_j^\top \left(\frac{\sum_{i=1}^g n_i c_i \mathbf{S}_i}{\sum_{i=1}^g n_i c_i^2} \right) \mathbf{A}_j \tag{5.45}$$

should be diagonal. This also indicates that PPCs and $\mathbf{A}$ can be found by consecutive EVD of $\frac{\sum_{i=1}^g n_i c_i \mathbf{S}_i}{\sum_{i=1}^g n_i c_i^2}$, where updated values for c_i and λ_j are found by (5.43) and (5.44). This is a very important feature which will be utilized in variable selection for dimension reduction.

Note that, as in the ML case, the equation for the proportionality constraints (5.43) holds also for $i = 1$, because $\sum_{j=1}^r \mathbf{a}_j^\top \mathbf{S}_1 \mathbf{a}_j \lambda_j = \sum_{j=1}^r \lambda_j^2$. Hence $c_1 = 1$.

The steps of the algorithm for solving the least square equation is outlined in Section 5.6.4, but we see from (5.43) to (5.45) that the LS estimates correspond much to what one would intuitively expect. For instance, the constants of proportionality ($c_i's$) are estimated as ratio of total squared variances (5.43). Alternatively, c_i can be estimated as the ratio of the total variations of two matrices. That is

$$c_i = \frac{\text{trace}(\mathbf{S}_i)}{\text{trace}(\mathbf{S}_1)}, \quad i = 2, \ldots, g, \tag{5.46}$$

where $\text{trace}(\mathbf{S}_i)$ is the total variation of the i^{th} group, which is given as

$$\text{trace}(\mathbf{S}_i) = \sum_{j=1}^{r} \lambda_{ij}. \tag{5.47}$$

5.6.2.1 Numerical Illustration

For Illustration we solve the PPC-LS problem for the Fisher's Iris data. The estimators are obtained by solving (5.38), making use of an alternative iterative

algorithm similar to the ML case.

$$\mathbf{A} = \begin{pmatrix} .7307 & -.2061 & .5981 & -.2566 \\ .2583 & .8568 & .1586 & .4171 \\ .6127 & -.2209 & -.6816 & .3336 \\ .1547 & .4178 & -.3906 & -.8056 \end{pmatrix}$$

and respectively:

$$\lambda_1^2 = \begin{pmatrix} 48.4509 \\ 6.2894 \\ 6.3261 \\ 1.4160 \end{pmatrix}, \lambda_2^2 = \begin{pmatrix} 69.2709 \\ 10.5674 \\ 5.2504 \\ 3.7482 \end{pmatrix}, \lambda_3^2 = \begin{pmatrix} 14.7542 \\ 7.9960 \\ 6.3983 \\ 1.7719 \end{pmatrix}.$$

The proportionality constants are estimated as $1.0000, 1.4284$ and $.3343$. For comparison with the ML solution obtained, we predict the estimated population covariance matrices for the Fisher's Iris Data:

$$\hat{\Sigma_1} = \begin{pmatrix} 28.0939 & 8.0037 & 19.7198 & 4.1039 \\ 8.0037 & 9.3609 & 5.9809 & 3.5642 \\ 19.7198 & 5.9809 & 21.1279 & 4.5960 \\ 4.1039 & 3.5642 & 4.5960 & 4.7767 \end{pmatrix}$$

$$\hat{\Sigma_2} = \begin{pmatrix} 40.1290 & 11.4325 & 28.1679 & 5.8621 \\ 11.4325 & 13.3711 & 8.5431 & 5.0911 \\ 28.1679 & 8.5431 & 30.1793 & 6.5650 \\ 5.8621 & 5.0911 & 6.5650 & 6.8231 \end{pmatrix}$$

$$\hat{\Sigma_3} = \begin{pmatrix} 9.3914 & 2.6755 & 6.5921 & 1.3719 \\ 2.6755 & 3.1292 & 1.9993 & 1.1915 \\ 6.5921 & 1.9993 & 7.0629 & 1.5364 \\ 1.3719 & 1.1915 & 1.5364 & 1.5968 \end{pmatrix}.$$

The value of the PPC-LS objective function is 129.1579. The fit achieved by the LS-CPC solution produced is 93.3166. In both examples we consider $n_i := n_i \diagup \sum_{i=1}^{g} n_i$.

5.6.3 Sparse discrimination using proportional CPC (SD-PCPC)

We have seen in Section 5.6.2 that we can estimate the parameters c_i, $\mathbf{A}$, and $\mathbf{\Lambda}_r$ by minimizing (5.38). However, we need to identify a small number of variables that are important for classification. The cardinality penalty was found to be effective in finding sparse common principal components. Therefore, as in SDCPC, we here also propose to impose the cardinality constraint on (5.38) to select a set of variables which have better classification performance as compared with other possible sets of variables. Thus the modified constrained minimization problem can be given as

$$\begin{aligned} \min_{\mathbf{a}} \left(\sum_{i=1}^{g} n_i ||\mathbf{S}_i - c_i \mathbf{A}\mathbf{\Lambda}_r \mathbf{A}^\top||_F^2 \right) \\ \text{s.t.} \quad \mathbf{A}^\top \mathbf{A} = \mathbf{I}_r, \ \mathbf{Card}(\mathbf{a}_k) \leq t, \end{aligned} \tag{5.48}$$

where the constraint $\mathbf{Card}(\mathbf{a}_k) \leq t$ means that the cardinality selects only t variables out of the original p variable from the k^{th} column of $\mathbf{A}$.

By letting, $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_r)$, the k^{th} vector $\mathbf{a}_k$, $k = 1, 2, \ldots, r$, can be sequentially found by solving the constrained minimization problem

$$\begin{aligned} &\min_{\mathbf{a}} \left(\sum_{i=1}^{g} n_i ||\mathbf{S}_i - c_i \mathbf{a}_k \mathbf{\Lambda}_r \mathbf{a}_k^\top||_F^2 \right) \\ &\text{s.t.} \quad \mathbf{a}_k^\top \mathbf{a}_k = 1, \quad \mathbf{a}^T \mathbf{A}_{k-1} = \mathbf{0}_{k-1}^T, \quad \mathbf{Card}(\mathbf{a}_k) \leq t. \end{aligned} \tag{5.49}$$

We have developed an algorithm that solves problem (5.49). The main steps of the SD-PCPC algorithm are summarized in Section 5.6.4 below.

5.6.4 Algorithm 4: SD-PCPC

1. Consider an $n \times p$ grouped multivariate data matrix.

2. Randomly split the data into two sets to form training and testing datasets. Let $\mathbf{X}$ denotes the training data set.

3. For cross-validation, randomly divide the training data into 10 subsets such that each subset contains one tenth of each group.

4. Take nine of the ten subsets and let $\mathbf{X}_{/m}$ denote the data set when the m^{th} subset is omitted and let $\mathbf{X}_c$ denote the omitted data.

5. Put $m = 1$.

6. For the data set $\mathbf{X}_{/m}$, find the covariance matrix for each group ($\mathbf{S}_i$), $i = 1, 2, \ldots g$.

7. For $i = 1, 2, \ldots, g$, put

$$c_i = \frac{\text{trace}(\mathbf{S}_i)}{\text{trace}(\mathbf{S}_1)}. \tag{5.50}$$

8. Start the cardinality with $t = 1$, where $t < p$.

9. For $k = 1, 2, \ldots r \leq \min(p, g-1)$, find the $p \times 1$ vector $\mathbf{a}_k$ by sequentially solving the problem.

$$\min_{\mathbf{a}} \left(\sum_{i=1}^{g} n_i ||\mathbf{S}_i - c_i \mathbf{a}_k \mathbf{\Lambda}_r \mathbf{a}_k^\top||_F^2 \right) \quad \text{s.t.} \quad \mathbf{a}_k^\top \mathbf{a}_k = 1, \;\; \mathbf{a}_k^T \mathbf{A}_{k-1} = \mathbf{0}_{k-1}^T, \;\; \mathbf{Card}(\mathbf{a}_k) \leq t. \tag{5.51}$$

10. Let the solutions of (5.51) be $\mathbf{a}_1^*, \mathbf{a}_2^*, \ldots \mathbf{a}_r^*$. Form a matrix $\mathbf{A}^* = (\mathbf{a}_1^*, \mathbf{a}_2^*, \ldots \mathbf{a}_r^*)$.

11. Classify the observations in the omitted data set, $\mathbf{X}_c$, using the classifiers $\mathbf{X}_c\mathbf{a}_1^*, \mathbf{X}_c\mathbf{a}_2^*, \ldots \mathbf{X}_c\mathbf{a}_r^*$. Record the number of misclassification, calling it Err(m,t).

12. Update t in the interval (1,20] if $p > 20$ and repeat steps 8-10.

13. If $m \leq 10$, increase m by 1 and repeat steps 5-10.

14. Find the value of t that minimizes $\sum_{m=1}^{10} Err(m, t)$. Using all the training data, repeat steps 6-9 for that value of t and let $\mathbf{a}_1$,$\mathbf{a}_2$, $\ldots \mathbf{a}_r$ be the solution to (5.51). The discriminant functions are $\mathbf{y}_1 = \mathbf{X}\mathbf{a}_1, \mathbf{y}_2 = \mathbf{X}\mathbf{a}_2, \ldots \mathbf{y}_r = \mathbf{X}\mathbf{a}_r$.

5.6.4.1 Notes on the algorithm

1. When we say, for example, that the first three principal components explain more than 80% of the total variation, the total variation is defined as the sum of the eigenvalues of the covariance matrix, which equals the trace of that matrix. In step 7 of the algorithm we use that definition of total variation to determine the c_i.

2. The procedure for evaluating the algorithm with real data is the same as for

algorithm 3. Thus, the performance of the resulting discriminant functions is evaluated on the test set.

5.6.5 Numerical illustration of SD-PCPC

To evaluate the performance of the SD-PCPC, we applied it to the six real data sets used earlier.

1. Fisher's iris data ($n > p$)
2. Rice data ($p > n$)
3. Ovarian Cancer data ($p >> n$)
4. Leukemia data ($p >> n$)
5. Ramaswamy data ($p >> n$)
6. IBD data ($p >> n$).

Table 5.3: *Constants of proportionality of sample covariance matrices of real data sets*

Data	g	c_1	c_2	c_3	c_4	c_5	c_6	$\cdots$	c_{14}
Iris	3	1.00	1.4284	0.3343	-	-	-	$\cdots$	-
Rice	4	1.00	0.8493	0.7695	0.6042	-	-	$\cdots$	-
Ovarian Cancer	2	1.00	0.4900	-	-	-	-	$\cdots$	-
Leukemia	6	1.00	0.7600	0.9900	1.1200	1.2700	0.8500	$\cdots$	-
IBD	3	1.00	0.4320	1.1567	-	-	-	$\cdots$	-
Ramaswamy	14	1.00	0.1300	0.03200	0.6300	0.0310	0.0112	$\cdots$	0.0333

It it assumed that $c_1 = 1$. The remaining c_i's, $i = 2, \ldots, g$ are given in Table 5.3. We can see that the group covariances matrices of the Iris, Rice, Leukemia, and IBD data sets vary comparatively little across groups appreciably in their total variance, and the group covariance matrices of Ramaswamy vary far more.

Using these c_i's , we further analysis the data sets using SD-PCPC and the summarized results are presented in Table 5.4.

Table 5.4: *Numerical results of SD-PCPC on low and high-dimensional real datasets*

Data	n	p	g	r	t	Error	Time
Iris	150	4	3	2	[2,2]	4%	0.0013
Rice	62	100	4	3	[3,3,3]	37.21%	0.0059
Ovarian Cancer	216	4,000	2	1	[10]	18.21%	21.0011
Leukemia	248	12,558	6	3	[14,14,14]	17.17%	68.01289
IBD	127	22,283	3	2	[13,13]	23.10%	155.3122
Ramaswamy	198	16,063	14	3	[13,13,13]	48.15%	139.1301

From Table 5.4, we can see that our new SD-PCPC performs well on the data sets Iris, Ovarian cancer, Leukemia, and IBD with misclassification rates 4%, 18.21%, 23.17%, and 23.10%, respectively. However, it performs weakly on the Rice and Ramaswamy data sets with misclassification rates 37.21% and 48.15%, respectively. The weak performance of the SD-PCPC on the rice data may be because of the tightness of the groups to each other (Krzanowski et al., 1995). Similarly, the weak performance of SD-PCPC on the Ramaswamy data set may be due to the fact that the Ramaswamy data set has many groups (i.e., g=14). Therefore, the SD-PCPC method does not seem to give better results than ran-

dom guessing when the number of groups is very large. However, in general, we conclude that the SD-PCPC performs well when the number of groups is fairly small.

5.7 Chapter summary

In this chapter, a sparse LDA based on CPC has been proposed for high dimensional classification problems. The sparse LDA with CPC (DSCPC) makes a weaker assumption than the assumption of equal group covariance matrices. This method is developed using the likelihood approach (Zou, 2006). The estimated CPCs are used as classification vectors. A cardinality penalty is used to achieve sparsity. This penalty helps to select a small variables from possibly a huge number of variables. From the numerical results using real data sets, sparse LDA based on CPC performs well. Furthermore, our newly proposed method is compared with two other existing methods using real data sets. Finally, we proposed that high-dimensional discrimination can also be performed using proportional CPCs when the group covariance matrices have some proportionality. We called the resulting sparse discrimination method SD-PCPC. SD-PCPC gives good results when group covariance matrices are approximately proportional to each other.

Chapter 6

Sparse LDA using optimal scoring

6.1 Introduction

As an alternative method for high-dimensional LDA, we propose a new method that uses optimal scoring (OS), called sparse LDA. The method is developed by using an l_1 minimization method and is commonly called the Dantzig selector (Candès and Tao, 2007) in statistical estimation when p is much larger than n. It assumes that in high dimensional discriminant analysis, most of the variables correspond to noise and only a few variables are important for classifying observations into their respective groups. Clemmensen et al. (2011) developed a sparse discriminant analysis based on OS but the algorithm has some convergence problems. Here, we aim to develop an effective sparse discriminant analysis using OS and the Dantzig selector.

Let us first define some notation for formulating the optimal scoring of discriminant analysis using the Dantzig selector. We recall that the multivariate data $\mathbf{X}$ consists of n observations, where each observation $\mathbf{x}_j$ comprises p-variables.

Let $\mathbf{Y}$ denote an $n \times g$ group indicator matrix, with columns that correspond to the dummy-variable codings of the g-groups. That is, $\mathbf{y}_{ij} \in \{0, 1\}$ indicates whether the j^{th} observation belongs to the i^{th} group. We assume that the columns of $\mathbf{X}$ are centered (i.e., orthogonal to the constant vector $\mathbf{1}$) so that the columns of $\mathbf{X}$ will have mean zero and the total sample covariance matrix will be $\mathbf{S} = n^{-1}\mathbf{X}^T\mathbf{X}$.

Our new method is called sparse linear discriminant analysis based on optimal scoring (SLDA-OS) that is developed based on the fact that discrimination problem can be recast as a regression problem. Using the same formulation as Dantzig selector, our discrimination method can be given as

$$\min ||\boldsymbol{\beta}_k||_1 \quad \text{subject to} \quad ||\mathbf{X}^T\mathbf{r}||_\infty \leq \lambda, \ \boldsymbol{\theta}_k^T\mathbf{Y}^T\mathbf{Y}\boldsymbol{\theta}_k = 1, \boldsymbol{\theta}_k^T\mathbf{Y}^T\mathbf{Y}\boldsymbol{\theta}_l = 0 \quad \text{for all} \quad l < k, \tag{6.1}$$

where $||.||_1$ and $||.||_\infty$ represent the l_1-norm and l_∞-norm, respectively, λ is a tuning parameter, and $\mathbf{r}$ is the vector of residuals given as:

$$\mathbf{r} = \mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k, \tag{6.2}$$

where $\boldsymbol{\theta}_k$ is a $g \times s$ matrix of scores, and $\boldsymbol{\beta}_k$ is a $p \times s$ matrix of regression coefficients. The theoretical and practical results of our method, SLDA-OS, will be given in the succeeding sections.

The chapter is organized as follows: It reviews the connection between multivariate regression analysis and discriminant analysis via optimal scoring in Section 6.2, and then the formulation of discrimination problem as regression problem is given in Section 6.3. We have proposed a new sparse LDA based on optimal scoring, SLDA-OS, in Section 6.4. This section shows the theoretical formu-

lation of discrimination problem as regression problem via optimal scoring, and the use of ℓ_1-minimization to select a small number of variables. The algorithm for SLDA-OS is given in Section 6.4.1. Section 6.5 presents numerical illustration of our method. The results of high-dimensional simulated and real data sets are given in this section. Finally, the summary of the chapter is given in Section 6.6.

6.2 Connection of multivariate regression analysis and discriminant analysis via optimal scoring

Without loss of generality, we assume that the columns of $\mathbf{X}$ have mean zero. Hastie et al. (1994) developed a multivariate regression procedure as a simpler way to perform classification. The regression procedure is applied to an indicator response $\mathbf{Y}$ that represents the classes, and a new observation is assigned to the class with the largest fitted value. This procedure was referred to as *softmax* (Hastie et al., 1994).

In the two-group case, with equal sample sizes, softmax is essentially equivalent to LDA. They may not be equivalent in general, but Hastie et al. (1994) showed that the space of LDA fits in the same space as the space in which multivariate linear regression fits. This means that the LDA solution can be obtained from a linear discriminant analysis of the fitted values from a multivariate regression. This equivalence was further proved and discussed in detail by Hastie et al. (1995). Using LDA in this fashion as a postprocessor for multivariate linear regression generally improves its classification performance.

Hastie et al. (1995) noted that discriminant variates are the same as the canon-

ical variates that result from a canonical correlation analysis (CCA), and they used the latter interchangeably with discriminant variates. It is less well known that an asymmetric version of canonical correlation analysis, called optimal scoring (OS), also yields a set of dimensions that coincide up to scalars with those of LDA and CCA.

Hastie et al. (1995) noted that OS, CCA and LDA are equivalent and showed the equivalence of the three methods when a penalization is imposed on each method for dimension reduction. Dimension reduction means reexpressing the data in fewer variables while minimizing the loss of essential information for the problem at hand. In discriminant analysis, such reduction can actually be beneficial when the "lost dimensions" show only spurious or weak structure. LDA based on OS is equivalent to CCA; the linear predictors define the one set of variables, and a set of dummy variables representing class membership defines the other set. CCA in this context gives the solution to a scoring problem that is described below.

Let $\mathbf{Y}$ be the $n \times g$ indicator matrix corresponding to the dummy-variable coding for the classes, with $y_{ij} = 1$ if the j^{th} observation belongs to the i^{th} group, and $y_{ij} = 0$ otherwise.

Let $\mathbf{\Theta}$ be a $g \times s$ matrix of scores, and $\mathbf{B}$ be a $p \times s$ matrix of regression coefficients, which are respectively given as:

$$\mathbf{\Theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_s) \tag{6.3}$$

where $\boldsymbol{\theta}_k$ is a $g \times 1$ vector, and

$$\mathbf{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_s) \tag{6.4}$$

where $\boldsymbol{\beta}_k$ is a $p \times 1$ vector, for $k = 1, 2, \ldots, s \leq \min(g-1, p)$.

Then the scores $\boldsymbol{\theta}_k$ and the coefficients $\boldsymbol{\beta}_k$ are chosen to minimize the problem:

$$\min\{||\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k||^2\}. \tag{6.5}$$

The scores are assumed to be mutually orthogonal and normalized with respect to an appropriate inner product to prevent trivial zero solutions.

If we let $\boldsymbol{\Theta}^*$ be the $n \times s$ matrix of transformed values of the classes, then it is clear that if the scores were fixed, we could minimize problem (6.5) by regressing $\boldsymbol{\Theta}^*$ on $\mathbf{x}$. Let $\mathbf{P}_X$ project onto the column space of the predictors. Then the scores are obtained by minimizing

$$\min \text{trace}\{\boldsymbol{\Theta}^{*T}(\mathbf{I} - \mathbf{P}_X)\boldsymbol{\Theta}^*\}/n \tag{6.6}$$

$$= \min \text{trace}\{\boldsymbol{\Theta}^T\mathbf{Y}^\top(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}\boldsymbol{\Theta}\}/n. \tag{6.7}$$

Hastie et al. (1995) developed an algorithm to solve problem (6.6). The steps of the algorithm are summarized as:

1. **Initialize**. Form $\mathbf{Y}$, the $n \times g$ indicator matrix corresponding to the dummy-variable coding for the classes.

2. **Multivariate regression**. Set $\hat{\mathbf{Y}} = \mathbf{P}_X\mathbf{Y}$ and denote the $p \times g$ coefficient matrix by $\mathbf{B}$: $\hat{\mathbf{Y}} = \mathbf{XB}$.

3. **Optimal scores**. Obtain the eigenvector matrix $\boldsymbol{\Theta}$ of $\mathbf{Y}^\top\hat{\mathbf{Y}} = \mathbf{Y}^\top\mathbf{P}_X\mathbf{Y}$ with normalization $\boldsymbol{\Theta}^\top\mathbf{D}\boldsymbol{\Theta} = \mathbf{I}$, where $\mathbf{D} = \mathbf{Y}^\top\mathbf{Y}/n$.

4. **Update**. Update the coefficient matrix in step 2 to reflect the optimal scores by setting $\mathbf{B} = \mathbf{B}\boldsymbol{\Theta}$. The final optimally scaled regression fit is the s vector function $\mathbf{B}^\top\mathbf{x}$.

There is an alternative algorithm for computing the usual canonical variates. The final coefficient matrix **B** is, up to a diagonal scale matrix, the same as the discriminant analysis coefficient matrix.

6.3 Linear discriminant analysis via optimal scoring

We recall from Chapters 2 and 3 that LDA can be considered as arising from Fisher's discriminant problem. Fisher's discriminant problem involves seeking discriminant vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_s$ that successively solve the problem

$$\max\{\boldsymbol{\beta}_k^\top \boldsymbol{\Sigma}_b \boldsymbol{\beta}_k\} \text{ subject to } \boldsymbol{\beta}_k^\top \boldsymbol{\Sigma}_w \boldsymbol{\beta}_l = \begin{cases} 1, & k = l, \\ 0, & k \neq l. \end{cases} \tag{6.8}$$

These solutions are directions found by maximizing the between-group variance relative to their within-group variance. However, for discrimination problem with $p > n$, the within-group covariance matrix has to be regularized to solve problem (6.8). For example, under the assumption that variables are independent, the within-group covariance matrix $(\boldsymbol{\Sigma}_w)$ can be replaced by its diagonal matrix. With this simplification we can solve problem (6.8) and find the discriminant vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_s$.

We can alternatively find $\boldsymbol{\beta}_k$'s using the formulation of discrimination problem via optimal scoring. Here, we assume that the discriminant analysis problem can be recast as a regression problem by changing categorical variables into quantitative variables via optimal scoring.

Let **Y** be the $n \times g$ indicator matrix corresponding to the dummy-variable

coding for the classes; that is, $y_{ij} = 1$ if the j^{th} observation belongs to the i^{th} group, and $y_{ij} = 0$ otherwise. Then the discrimination problem using optimal scoring has the form

$$\min\{||\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k||^2\} \text{ subject to } \frac{1}{n}\boldsymbol{\theta}_k^\top \mathbf{Y}^\top \mathbf{Y}\boldsymbol{\theta}_l = \begin{cases} 1, & k = l, \\ 0, & k \neq l, \end{cases} \tag{6.9}$$

where $\boldsymbol{\theta}_k$ is a $g \times 1$ vector of scores, and $\boldsymbol{\beta}_k$ is a $p \times 1$ vector of coefficients, for $k = 1, 2, \ldots, s \leq \min(g-1, p)$, and $||.||$ denotes the vector ℓ_2-norm defined by $\sqrt{y_1^2 + y_2^2 + \cdots + y_n^2}$ for all $\mathbf{y} \in \Re^n$. If we let $\mathbf{D} = \frac{1}{n}\mathbf{Y}^\top \mathbf{Y}$ be a diagonal matrix of group proportions, the constrains in (6.9) can be redefined as $\boldsymbol{\theta}_k^\top \mathbf{D}\boldsymbol{\theta}_k = 1$ and $\boldsymbol{\theta}_k^\top \mathbf{D}\boldsymbol{\theta}_l = 0$ for $k \neq l$. The vector $\boldsymbol{\beta}_k$ that solves (6.9) is proportional to the solution to (6.8) (Clemmensen et al., 2011). We will refer to the vector that solves (6.9) as the k^{th} discriminant vector. Performing LDA on $\mathbf{X}$ yields the s classifiers $\mathbf{X}\boldsymbol{\beta}_1, \ldots, \mathbf{X}\boldsymbol{\beta}_s$.

For classification problem with $p >> n$ data, Clemmensen et al. (2011) proposed a variant method of sparse discriminant analysis based on the optimal scoring problem that employs regularization via the elastic net penalty function. Suppose we have identified the first $k-1$ discriminant vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{k-1}$ and scoring vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-1}$. Then the k^{th} sparse discriminant vector $\boldsymbol{\beta}_k$ and scoring vector $\boldsymbol{\theta}_k$ are found as the optimal solutions to the optimal scoring criterion

problem

$$\min\{||\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k||^2 + \gamma\boldsymbol{\beta}^\top\Omega\boldsymbol{\beta} + \lambda||\boldsymbol{\beta}||_1\} \text{ subject to } \frac{1}{n}\boldsymbol{\theta}_k^\top\mathbf{Y}^\top\mathbf{Y}\boldsymbol{\theta}_l = \begin{cases} 1, & k = l, \\ 0, & l < k, \end{cases} \tag{6.10}$$

where γ and λ are nonnegative tuning parameters and Ω is a $p \times p$ positive definite matrix. The optimization problem (6.10) is nonconvex, due to the presence of nonconvex spherical constraints. Consequently, it may not converge to a globally optimal solution using iterative procedures. Moreover, it is computationally very expensive, especially when both p and m are very large, where m is the number of nonzero coefficients.

Our primary objective is to develop an alternative sparse discrimination problem via optimal scoring. But we still keep the assumption that a discrimination problem can be recast as a regression problem. We formulate our new sparse LDA with optimal scoring in a similar fashion used with the Danztig selector in regression analysis for $p > n$.

6.4 Sparse LDA using optimal scoring

Our aim is to develop an efficient method of discrimination based on optimal scoring. We have reviewed various methods of discriminant analysis for high dimensional classification problem in Chapter 3. We have also briefly reviewed two relevant methods in Section 6.3 above. We observe that there is still a need to develop an alternative method of discrimination based on optimal scoring that improves the weakness of the exiting methods. We are now propose that sparse

discrimination can be achieved by adapting the Dantzig selector to the discrimination problem. The Dantzig selector was found to be an efficient method in regression analysis when $p >> n$. Hence, we propose in this chapter that high-dimensional discriminant analysis can be alternatively solved using the Dantzig selector. First, let us briefly review the Dantzig selector in regression analysis.

The Dantzig selector (Candès and Tao, 2007) has already received a considerable amount of attention. It was defined for linear regression model where $p > n$ and the set of coefficients is sparse, i.e, most of the β's are 0. The k^{th} Dantzig estimate $\hat{\boldsymbol{\beta}}_k$ is defined as the solution to

$$\min ||\hat{\boldsymbol{\beta}}_k||_1 \quad \text{subject to} \quad ||\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_k)||_\infty \leq \lambda, \tag{6.11}$$

where $||.||_1$ and $||.||_\infty$ represent the ℓ_1- and ℓ_∞-norms,respectively and λ is a tuning parameter. Candès and Tao (2007) gave detailed theoretical and practical results to substantiate that regression coefficient vector $\hat{\boldsymbol{\beta}}_k$ that solves (6.11) is a very effective estimate in regression problems with $p >> n$.

By adopting the formulation of Danzig selector (6.11) and by using notation from Section 6.3, and imposing appropriate constraint, we define our sparse LDA using the optimal scoring (SLDA-OS) problem as:

$$\min ||\hat{\boldsymbol{\beta}}_k||_1 \quad \text{subject to} \quad ||\mathbf{X}^\top(\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k)||_\infty \leq \lambda, \tag{6.12}$$

$$\text{and} \quad \frac{1}{n}\boldsymbol{\theta}_k^\top \mathbf{Y}^\top \mathbf{Y}\boldsymbol{\theta}_l = \begin{cases} 1, & k = l, \\ 0, & k \neq l. \end{cases}$$

As before, $\boldsymbol{\theta}_k$ is a $g \times 1$ vector of scores. By letting $\mathbf{D} = \frac{1}{n}\mathbf{Y}^\top\mathbf{Y}$, the constrains in

(6.12) can be rewritten as $\boldsymbol{\theta}_k^\top \mathbf{D}\boldsymbol{\theta}_k = 1$ and $\boldsymbol{\theta}_k^\top \mathbf{D}\boldsymbol{\theta}_l = 0$ for $k \neq l$. We refer to the $\boldsymbol{\beta}_k$ that solves (6.12) as the k^{th} discriminant vector.

We use an iterative algorithm to solve (6.12) and adapt a similar procedure to that used by Clemmensen et al. (2011) to solve problem (6.10). That is, the algorithm involves holding $\boldsymbol{\theta}_k$ fixed and optimizing with respect to $\boldsymbol{\beta}_k$, then holding $\boldsymbol{\beta}_k$ fixed and optimizing with respect to $\boldsymbol{\theta}_k$, repeating this until convergence. For fixed $\boldsymbol{\theta}_k$, we obtain

$$\min ||\hat{\boldsymbol{\beta}}_k||_1 \quad \text{subject to} \quad ||\mathbf{X}^\top(\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k)||_\infty \leq \lambda. \tag{6.13}$$

Problem (6.13) is exactly the same as the Dantzig selector except we use $\mathbf{Y}\boldsymbol{\theta}_k$ as a response variable instead of just $\mathbf{Y}$. Therefore, problem (6.13) can be solved using the Danzig selector algorithm. For fixed $\boldsymbol{\beta}_k$, the optimal scores $\boldsymbol{\theta}_k$ solve the problem

$$\begin{aligned} &\min ||\hat{\boldsymbol{\beta}}_k||_1 \quad \text{subject to} \quad ||\mathbf{X}^\top(\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k)||_\infty \leq \lambda, \\ &\text{and} \quad \boldsymbol{\theta}_k^\top \mathbf{D}\boldsymbol{\theta}_k = 1, \quad \boldsymbol{\theta}_k^\top \mathbf{D}\boldsymbol{\theta}_l = 0 \quad \text{for} \quad k \neq l. \end{aligned} \tag{6.14}$$

Problem (6.14) can be solved by modifying the SDA algorithm (Clemmensen et al., 2011). Let $\mathbf{Q}_k$ be the $g \times k$ matrix consisting of the previous $k-1$ solutions of $\boldsymbol{\theta}_k$, as well as the trivial solution vector of all 1s. We can show that the solution to (6.14) is given by $\boldsymbol{\theta}_k = c(\mathbf{I} - \mathbf{Q}_k\mathbf{Q}_k^\top \mathbf{D})\mathbf{D}^{-1}\mathbf{Y}^\top \mathbf{X}\boldsymbol{\beta}_k$, where c is a proportionality constant such that $\boldsymbol{\theta}_k^\top \mathbf{D}\boldsymbol{\theta}_k = 1$. $\mathbf{D}^{-1}\mathbf{Y}^\top \mathbf{X}\boldsymbol{\beta}_k$ is the unconstrained estimate for $\boldsymbol{\theta}_k$, and the term $(\mathbf{I} - \mathbf{Q}_k\mathbf{Q}_k^\top \mathbf{D})$ is the orthogonal projector (in $\mathbf{D}$) onto the subspace of R^k orthogonal to $\mathbf{Q}_k$.

Let $\mathbf{r} = \mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k$. There are two reasons why the size of the correlated residual vector $\mathbf{X}^T\mathbf{r}$ is constrained rather than the residual vector $\mathbf{r}$. The first

reason is that because of the invariance property, i.e, the estimation procedure (6.14) is invariant with respect to orthogonal transformation applied to the data vector since the feasible region is invariant. The other reason is that the optimal program (6.14) is convex and it can easily be recast as a linear program (LP),

$$\min \sum_i u_i \quad \text{subject to} \quad -\mathbf{u} \leq \boldsymbol{\beta} \leq \mathbf{u}, \quad \text{and}$$
$$-\lambda \mathbf{1} \leq \mathbf{X}^T(\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k) \leq \lambda \mathbf{1} \tag{6.15}$$

where $\mathbf{u}$ and $\boldsymbol{\beta}_k$ are the optimization variables, and $\mathbf{1}$ is a p-dimensional vector of ones. Therefore, the estimation procedure is computationally feasible.

However, the constraint $||\mathbf{X}^\top(\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k)||_\infty \leq \lambda$ in problems (6.10) to (6.14) needs to be redefined. We note that the lower bound of $||\mathbf{X}^\top(\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k)||_\infty$ cannot, in general, be exactly zero. Since $\mathbf{Y}\boldsymbol{\theta}_k \neq \mathbf{X}\boldsymbol{\beta}_k$, there may be a situation where we cannot find a solution under the constraint $||\mathbf{X}^\top(\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k)||_\infty \leq \lambda$. Therefore, we must improve the constraint so as to get a solution all the time. One possible way of avoiding the nonexistence of a solution is to use the constraint

$$\left\{ ||\mathbf{X}^\top(\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k)||_\infty - ||\mathbf{X}^\top(\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\hat{\boldsymbol{\beta}}_k)||_\infty \right\} \leq \lambda, \tag{6.16}$$

where $\hat{\boldsymbol{\beta}}_k$ minimizes $||\mathbf{X}^\top(\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\hat{\boldsymbol{\beta}}_k)||_\infty$.

By using the constraint (6.16) in problem (6.14), our SLDA-OS problem becomes

$$\min ||\hat{\boldsymbol{\beta}}_k||_1 \quad \text{subject to} \quad \left\{ ||\mathbf{X}^\top(\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\boldsymbol{\beta}_k)||_\infty - ||\mathbf{X}^\top(\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\hat{\boldsymbol{\beta}}_k)||_\infty \right\} \leq \lambda, \tag{6.17}$$
$$\text{and} \quad \boldsymbol{\theta}_k^\top \mathbf{D}\boldsymbol{\theta}_k = 1, \quad \boldsymbol{\theta}_k^\top \mathbf{D}\boldsymbol{\theta}_l = 0 \quad \text{for} \quad k \neq l.$$

where $\hat{\boldsymbol{\beta}}_k$ minimizes $||\mathbf{X}^\top(\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}\hat{\boldsymbol{\beta}}_k)||_\infty$. Now we can find a solution for problem (6.17) using a small nonnegative value of the tuning parameter λ. The value of λ is found using a 10-fold cross validation given in the algorithm in Section 6.4.1. Moreover, problem (6.17) gives sparse discriminant vectors $\boldsymbol{\beta}_k$, because the ℓ_1-norm of $||\boldsymbol{\beta}_k)||_1$ defined as $\min(||\boldsymbol{\beta}_k)||_1) = \min(|\beta_1| + |\beta_2| + \cdots + |\beta_p|)$ makes some of the β's exactly zero.

The l_1-minimization produces coefficient estimates that are exactly 0 in a similar fashion to the Lasso and hence can be used as a variable selection method (James et al., 2009).

This minimization method leads to the sparsest solution over all feasible solutions (Candès and Tao, 2007). In other words, the objective is to find an estimator $\boldsymbol{\beta}_k$ with minimum number of nonzero components (as measured by the l_1-norm) among all objects that are consistent with the data. As the constraint on the residual vector is relaxed, the solution becomes more sparse.

(Candès and Tao, 2007) suggested using $\lambda = \sqrt{2\log p}$, which is equal to $\sqrt{2\log n}$ in the orthogonal design setting. Under this setting, the oracle properties of the Dantzig selector are in line with shrinkage results that are assumed to be optimal in the minimax sense. Furthermore, it will be interesting to find an optimal regularization factor using different methods such as cross-validation.

The goal in developing this method is to find the sparsest solution for (6.17). (Candès and Tao, 2007) have shown that the Dantzig selector produces the sparsest solution under the UUP condition. The UUP condition roughly states that for any small set of predictors, the s-vectors are nearly orthogonal to each other.

Moreover, due to the nature of linear programming, the problem in (6.17) can be solved quickly and efficiently. Consequently, the Dantzig selector is usually faster to implement than other existing methods, such as the Lasso (Candès and Tao, 2007). Another study by James et al. (2009) has shown that the Lasso and the Dantzig selector have connections. However, when the corresponding solutions are not identical, the Dantzig selector seems to give sparser solution than the lasso.

In general, we hope that the sparse LDA by optimal scoring based on the Dantzig selector will achieve the following objectives:

- to produce sparse and interpretable discriminant vectors in high-dimensional settings;
- to minimize computational cost.

We have developed an iterative algorithm to solve problem (6.17). The main steps of the algorithm are given in Section (6.4.1) below.

6.4.1 Algorithm 5: SLDA-OS

The main steps of the SLDA-OS algorithm are the following.

1. Let $\mathbf{X}$ be an $n \times p$ grouped multivariate data matrix and assume that $\mathbf{X}$ has been centered so that the columns of $\mathbf{X}$ have mean zero.
2. Form $\mathbf{Y}$, an $n \times g$ indicator matrix corresponding to the dummy-variable coding for the groups, defined by $y_{ij} = 1$ if the j^{th} observation belongs to the i^{th} group, and $y_{ij} = 0$ otherwise.

3. Form a full matrix, $\mathbf{T} = (\mathbf{X}, \mathbf{Y})$. Randomly split $\mathbf{T}$ into two sets to form training and testing datasets. Let $\mathbf{T}_1 = (\mathbf{X}_1, \mathbf{Y}_1)$ and $\mathbf{T}_2 = (\mathbf{X}_2, \mathbf{Y}_2)$ denote the training and testing data sets, respectively.

4. For cross-validation, divide randomly $\mathbf{T}_1$ into 10 subsets such that each subset contains one tenth of each group. Take nine of the ten subsets $\mathbf{T}_1$. Let $\mathbf{X}_{/m}$ and $\mathbf{Y}_{/m}$ denote the data sets of $\mathbf{T}_1$ when the m^{th} subset is omitted and let $\mathbf{X}_c$ and $\mathbf{Y}_c$ denote the omitted data of $\mathbf{T}_1$.

5. Put m=1.

6. Let $\mathbf{D} = \frac{1}{n^*}\mathbf{Y}_{/m}^\top \mathbf{Y}_{/m}$, where n^* is the number of observations in $(\mathbf{X}_{/m}, \mathbf{Y}_{/m})$.

7. Let $\mathbf{Q}_k$ be a $g \times k$ matrix consisting of the previous $k-1$ solutions $\boldsymbol{\theta}_k$. Start with $\mathbf{Q}_1$ as a matrix of 1's.

8. Start the tuning parameter, λ, with a small positive number.

9. For $k = 1, 2, \ldots, s \leq \min(g-1, p)$, compute the k^{th} discriminant solution pair $(\boldsymbol{\theta}_k, \boldsymbol{\beta}_k)$ as follows:

   (a) Initialize $\boldsymbol{\theta}_k = (\mathbf{I} - \mathbf{Q}_k\mathbf{Q}_k^\top \mathbf{D})\boldsymbol{\theta}_*$, where $\boldsymbol{\theta}_*$ is a random g-vector, and then normalize $\boldsymbol{\theta}_k$ so that $\boldsymbol{\theta}_k^\top \mathbf{D}\boldsymbol{\theta}_k = 1$.

   (b) For $t = 0, 1, 2 \ldots T$ until convergence or until a maximum iteration is reached, let $\boldsymbol{\beta}_k$ be the solution of the ι_1- minimization problem

$$\min_{\boldsymbol{\theta}_k, \boldsymbol{\beta}_k} ||\boldsymbol{\beta}_k||_1 \quad \text{s.t.} \quad \left\{ ||\mathbf{X}_{/m}^\top (\mathbf{Y}_{/m}\boldsymbol{\theta}_k - \mathbf{X}_{/m}\boldsymbol{\beta}_k)||_\infty - ||\mathbf{X}_{/m}^\top (\mathbf{Y}_{/m}\boldsymbol{\theta}_k - \mathbf{X}_{/m}\hat{\boldsymbol{\beta}}_k)||_\infty \right\} \leq \lambda, \tag{6.18}$$

   where $\hat{\boldsymbol{\beta}}_k$ minimizes $||\mathbf{X}_{/m}^\top (\mathbf{Y}\boldsymbol{\theta}_k - \mathbf{X}_{/m}\hat{\boldsymbol{\beta}}_k)||_\infty$.

(c) For fixed $\boldsymbol{\beta}_k$, update $\boldsymbol{\theta}_k$ as

$$\boldsymbol{\theta}_k = \frac{\mathbf{w}}{\sqrt{\mathbf{w}^\top \mathbf{D}\mathbf{W}}}, \text{ where } \mathbf{w} = (\mathbf{I} - \mathbf{Q}_k\mathbf{Q}_k^\top\mathbf{D})\mathbf{D}^{-1}\mathbf{Y}_{/m}^\top\mathbf{X}_{/m}\boldsymbol{\beta}_k.$$

10. If $k < s$, set $\mathbf{Q}_{k+1} = (\mathbf{Q}_k : \boldsymbol{\theta}_k)$.

11. Classify the observations in the omitted data set $(\mathbf{X}_c, \mathbf{Y}_c)$ using $\mathbf{X}_c\boldsymbol{\beta}_k$ as the classifier. Record the number of misclassifications, calling it Err(m,λ).

12. Change λ and repeat steps 9-11 until the full range of values of λ of interest has been considered.

13. If $m \leq 10$, increase m by one and repeat steps 6-12.

14. Find the value of λ that minimizes $\sum_{m=1}^{10} \text{Err}(m, \lambda)$. Using all the data, repeat steps 6-10 for that value of λ to obtain the optimal discriminant vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_s$.

15. Classification is performed using the usual classification rule of standard LDA. That is, we compute $\mathbf{X}\boldsymbol{\beta}_1, \mathbf{X}\boldsymbol{\beta}_2, \ldots, \mathbf{X}\boldsymbol{\beta}_s$ and assign each observation to its nearest centroid in this transformed space.

16. The performance of the resulting discriminant functions is evaluated on the test data set ($\mathbf{T}_2$).

6.5 Numerical illustration

We applied the new SLDA-OS algorithm to both simulated and real data sets.

6.5.1 Application to simulated data

We generated three data sets with different settings. The three simulated data sets were generated as follows:

Model 1: There are two groups of multivariate normal distributions, N($\boldsymbol{\mu}_1, \boldsymbol{\Sigma}$) and N($\boldsymbol{\mu}_2, \boldsymbol{\Sigma}$), each of dimension $p = 10,000$. The components of $\boldsymbol{\mu}_1$ are assumed to be 0 and for $\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_{2j} = 0.6$ if $j \leq 200$ and 0 otherwise. The covariance matrix $\boldsymbol{\Sigma}$ is the block diagonal matrix with ten blocks of dimension 1000×1000 whose element (j,j') is $0.6^{|j-j'|}$. For each class 100 training samples and 50 testing samples were generated. (i.e., n=300, p=10,000, g=2).

Model 2: There are three groups each assumed to have a multivariate normal distribution N($\boldsymbol{\mu}_i, \boldsymbol{\Sigma}$), $i = 1, 2, 3$ with dimension $p = 10,000$. The first 35 components of $\boldsymbol{\mu}_1$ are 0.7, $\boldsymbol{\mu}_{2j} = 0.6$ if $36 \leq j \leq 70$ and $\boldsymbol{\mu}_{3j} = 0.7$ if $71 \leq j \leq 105$ and 0 otherwise. All elements on the main diagonal of the covariance matrix $\boldsymbol{\Sigma}$ are equal to 1 and all other are equal to 0.6. For each class, we generated 100 training samples and 50 testing samples. (i.e., n= 450, p=10,000, g=3).

Model 3: There are three groups that were generated as: for $l \in \pi_i$ then $X_{lj} \sim N((i-1)/2, 1)$ if $j \leq 100$, $i = 1, 2, 3$ and $X_{lj} \sim N(0, 1)$ otherwise with dimension $p = 10,000$. A total of 200 training samples and 100 testing samples are generated. (i.e., n=300, p=10,000, g=3).

We employed cross-validation to choose the tuning parameter λ. We applied our method to the three simulated data sets, and compared it with another existing method, SDA. The results of the analysis are summarized in Table 6.1. Sparsity in Table 6.1 denotes the percentage of nonzero components.

Table 6.1: *Misclassification rate (in %), time (in seconds), and sparsity (in %) of two methods on the testing sets of three simulated data sets.*

Model	SLDA-OS			SDA		
	Error	Time	Sparsity	Error	Time	Sparsity
Model 1	4.50	12.50	11.33	13.0	12.50	21
Model 2	13.22	14.03	15	13.21	14.00	25.65
Model 3	12.11	13.67	14.5	14.80	12.50	31.60

The results in table show that our method (SLDA-OS) performed better than SDA for the first and third models. That is, SLDA-OS gave lower misclassification errors than SDA in models 1 and 3. The performance of both methods is almost the same for the second model. Moreover, the two methods were also compared based on their speed, and it was found that there is no significant difference between the speeds of the two methods. But, the SLDA-OS gave sparser discriminant vectors than SDA, as shown by the percentage of nonzero components in Table 6.1.

6.5.2 Application to real data sets

To further evaluate the performance of SLDA-OS, we applied it to the six real data sets that were used in Chapters 4 and 5. The six real data sets are

1. Fisher's iris data ($n > p$)

2. Rice data ($p > n$)

3. Ovarian Cancer data ($p >> n$)

4. Leukemia data ($p >> n$)

5. Ramaswamy data ($p >> n$)

6. IBD data ($p >> n$).

We analysed the data sets using SLDA-OS and the summarized results are presented in Table 6.2. We also included the results of two existing methods, SDA (Clemmensen et al., 2011) and PLDA (Witten and Tibshirani, 2011) for comparison.

Table 6.2: *Misclassification rate (in %) and time (in seconds) of three sparse LDA methods on the testing sets of six real data sets.*

Data	SLDA-OS		SDA		PLDA	
	Error	Time	Error	Time	Error	Time
Iris	3.00	0.0013	3.0	0.0013	4.00	0.0120
Rice	36.20	0.0068	37.15	0.0070	38.00	0.0760
IBD	30.00	121.0200	30.65	112.2230	34.50	131.0600
Leukemia	21.50	19.6289	27.65	19.9700	27.33	35.2000
Ovarian Cancer	5.10	55.31280	19.31	58.3452	20.65	60.1024
Ramaswamy	16.33	113.1340	16.16	116.5012	–	–

We can see from Table 6.2 that our new method SLDA-OS performs better than the other existing methods on data sets Rice, IBD, Leukemia, and Ovarian cancer, with misclassification rates (in %) 36.20, 30.00, 21.50, and 5.10 respectively. It also performs as well as SDA on the Iris Fisher's data set with a misclassifica-

tion rate of 3%, which is lower than the misclassification rate of PLDA. Further, for the Ramaswamy data, our method performs with a misclassification rate of 16.33%, which is very close to the performance of SDA. A noticeable features of the results for our method is that it performs classification of the Ovarian cancer data with only a 5.10% misclassification rate, far lower than with the other competing methods. We know that the ovarian cancer data set is a two-group data set. Hence, it seems that our new method can be very effective in classifying observations in a binary-group classification problem.

Regarding sparsity, on average for all data sets the SLDA-OS selected 20.18% of the variables while the SDA and PLDA selected 21.35% and 40.65% of the variables, respectively, to achieve the classification rates given in Table 6.2. Hence, the discriminant vector obtained by SLDA-OS has only about 20% nonzero components, which is similar to the most sparse of the other methods. Interpretation is much improved as only a small number of variables were selected from the original large number.

The tuning parameter λ was selected using a 10-fold cross validation. For illustration, we report the results from cross validation as λ varies for the Ovarian Cancer and Ramaswamy data sets. The cross validation results are presented in Figure 6.1 for Ovarian cancer data set, and in Figure 6.2 for Ramaswamy data set. We can see Figure 6.1 that the misclassification rate (MCE) decreases steadily until it reaches its minimum and stabilizes in the interval $\lambda \in (0.0015, 0.0027)$ before it starts rising again. So we can choose any value of λ in that interval, we selected $\hat{\lambda} = 0.002$. This gave the smallest misclassification rate of 5.1% for classification of

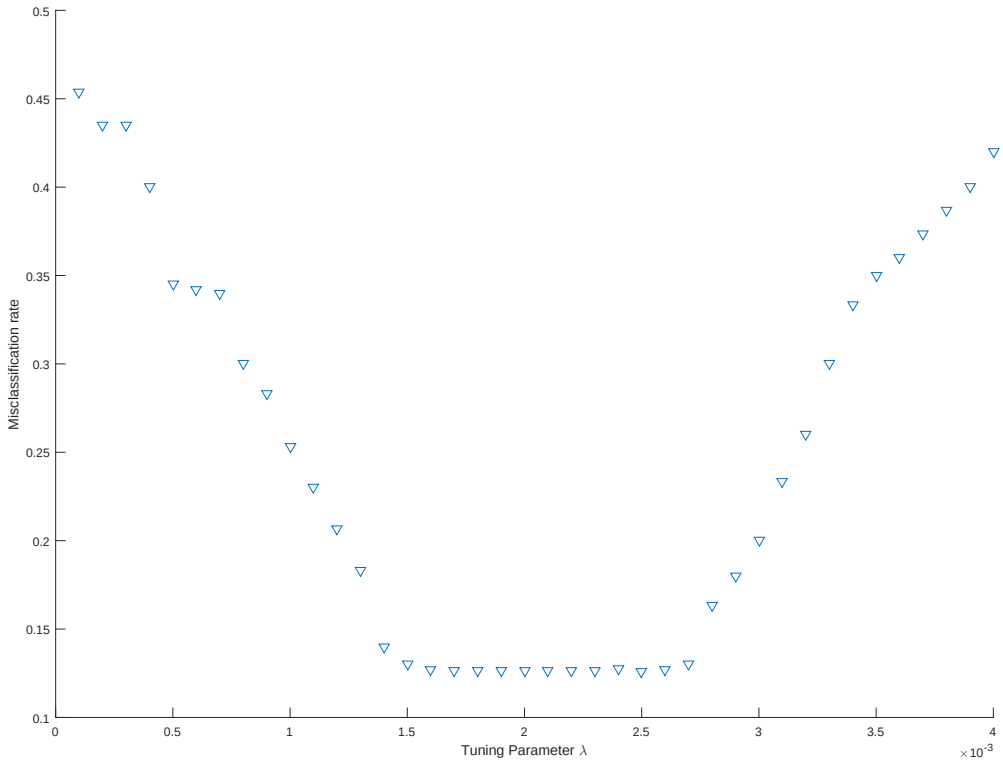

Figure 6.1: *The misclassification rate of the training set of the ovarian cancer data for different values of the tuning parameter (λ) resulting from cross-validation of SLDA-OS method.*

the Ovarian cancer data. Similarly, Figure 6.2 illustrates that the MCE decreased until it attained its minimum in the same interval of λ as for classification of the Ramaswamy data. Thus, in this case also, we again selected $\hat{\lambda} = 0.002$ though this gave a comparatively poor MCE of 16.23%.

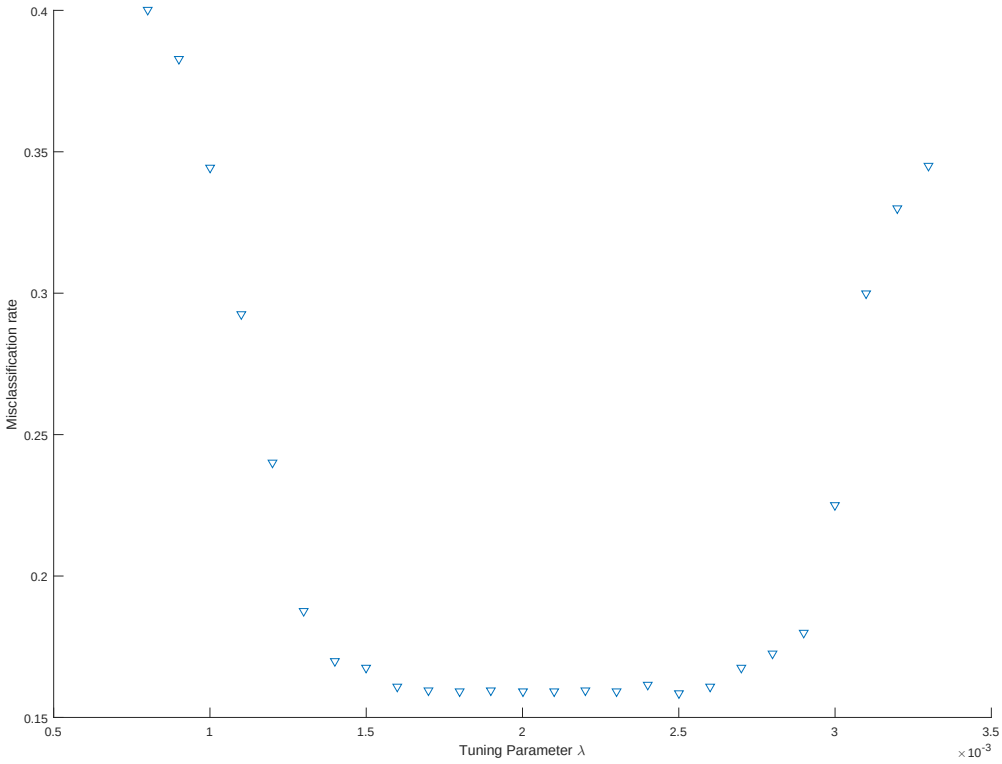

Figure 6.2: *The misclassification rate of the training set of the Ramaswamy data for different values of the tuning parameter λ resulted from cross-validation of SLDA-OS method.*

6.6 Chapter summary

The traditional LDA fails classifying observations in to their groups if the number of variables is very large relative to the number of observations. In this chapter, we propose an alternative sparse LDA for high-dimensional discrimination problem. Our proposal is based on the fact that discrimination problem can be recast as a regression problem via optimal scoring. Thus we call our method sparse LDA based on optimal scoring (SLDA-OS). Our approach extended the

LDA to the high-dimensional setting in such a way that the resulting discriminant vectors involve only a small number of the variables. The formulation of our method has a similar form with the Danzig selector and we employed the ℓ_1-minimization penalty to achieve the required sparsity.

We applied our method, SLDA-OS, to both simulated and real data sets with $p >> n$. It gives better results than other existing methods in terms of classification accuracy and speed. Most notably, our algorithm was found superior in binary classification to the two existing methods PLDA (Witten and Tibshirani, 2011) and SDA Clemmensen et al. (2011). In general, our sparse discriminant analysis method based on the Dantzig selector gives interpretable discriminant functions with relatively lower classification error and smaller number of nonzero variables. Hence, this method can be considered as a better alternative discrimination method when $p >> n$.

Chapter 7

General conclusions and future research

linear discriminant analysis is a method of identifying linear combinations of variables, called linear discriminant functions, that separates two or more groups and is useful for classifying items into groups. However, the traditional discriminant analysis is not applicable when the number of variables is greater than the number of observations. This thesis deals with LDA methods that can be applied to high-dimensional classification problems, where the number of variables is greater than the number of observations, and focuses on methods that give sparse discriminant functions, as this gives more interpretable classifiers.

7.1 Summary and conclusions

Chapter 2 briefly introduced the general discriminant analysis framework and presented various techniques of classical discriminant analysis to give a gen-

eral background. Three different approaches to discriminant analysis were presented and it was seen that most of the existing high-dimensional discriminant analysis methods use the classical methods as a basis for their development. That is, the high dimensional discriminant methods are the extension of classical discrimination methods obtained by modifying or improving the original formulations.

When the number of variables (p) is much larger than the number of observations (n), commonly written as $p >> n$, the classical linear discriminant analysis (LDA) does not perform classification effectively for three major reasons. First, the sample covariance matrix is singular and cannot be inverted. Second, high-dimensionality makes direct matrix operation very difficult if not impossible, hence hindering the applicability of the traditional LDA method. Although we may use the generalized inverse of the covariance matrix, the estimate is highly biased and unstable and will generally lead to a classifier with poor performance due to lack of observations. Also, computing eigenvalues of a large matrix can be challenging. Third, in the $p >> n$ scenarios when p is extremely very large, it is not only computationally difficult to find the discriminant functions but also interpretation is a serious problem. That is, we cannot identify which set of variables are accountable for classifying an observation into its right group. However, some methods have been proposed to tackle these difficulties as we reviewed in Chapter 3.

In Chapter 3, we reviewed some of the existing discriminant approaches that have been developed for in the high-dimensional setting. The chapter reviewed

approaches that emphasise dimension reduction in Section 3.1 and regularization in Section 3.2. Many of them used dimension reduction methods such as PCA or variable selection methods in a separated step before classification. Different models for dimension reduction that were given in Table 3.1.

Other methods that were reviewed in Section 3.2.1 that assume the variables in high-dimensions are independent. These methods use the independence assumption merely to overcome the problem of singularity, regardless of the accuracy of classification. The independence methods were developed based on the models 3-5 that are given in Table 3.1. Though these methods are computationally attractive, they do not involve the idea of sparsity or aim to produce interpretable results. Moreover, other groups of methods reviewed in Chapter 3 use regularized $\mathbf{W}$. The solution based on regularization may ease computational difficulty, but it gives less attention to variable selection (i.e. sparsity) which is a basic requirement in dealing with high dimensional discriminant analysis. In addition, all regularization methods require tuning a parameter which may not be easy unless cross-validation is used appropriately. Another drawback of several of the reviewed methods is that they deal with classification problems involving only two groups.

Therefore, we have proposed 5 alternative sparse discrimination methods that are given in Chapters 4-6 to fill the gap that still exists in high-dimensional classification problems. The 5 methods were developed based on various assumptions of group covariance matrices. We give the various assumptions that we used to develop our methods in Table 7.1. These 5 methods were applied to 6 selected

real data sets and the summarized results of all our methods and two other exiting methods are given in Table 7.2. We summarize and discuss the theoretical backgrounds and practical results of our methods below.

Table 7.1: *Assumptions about covariance matrices made by the five methods proposed in this thesis.*

Method	Assumptions
FC-SCLDA	$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$=diag$(\sigma_1^2, \ldots, \sigma_2^2)$
FC-SLDA2	$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$=diag$(\sigma_1^2, \ldots, \sigma_2^2)$ with $\lambda_d = 0$
SDCPC	$\boldsymbol{\Sigma}_i = \mathbf{A}\boldsymbol{\Lambda}_i\mathbf{A}^\top$
SD-PCPC	$\boldsymbol{\Sigma}_i = c_i\boldsymbol{\Sigma}_1$
SLDA-OS	$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$

In Chapter 4, we have proposed an alternative method called Function-constrained sparse LDA (FC-SLDA) and its simplified version (FC-SLDA2) for high-dimensional discriminant analysis. The constrained ℓ_1-minimization penalty is imposed on the discrimination problem to achieve sparsity. The ℓ_1-minimization is a popular technique in regression analysis to select variables when $p >> n$. For example, Candès and Tao (2007) used the Dantzig selector for selecting variables in regression analysis with $p >> n$ using the ℓ_1-minimization penalty.

FC-SLDA is developed based on Model-4 in Table 3.1. That is, it assumes that all group covariance matrices are equal and the common within-group covariance matrix is diagonal. Consequently, we used the diagonal within-group covariance $\mathbf{W}_d$ to circumvent the singularity problem. This is because an esti-

mate of $\mathbf{W}^{-1}$ does not necessarily provide a better classifier. For example, Fan et al. (2008) showed that the LDA can not be better than random guessing when the number of variables is larger than the sample size due to noise accumulation in estimating the covariance matrix. Another method developed by Witten and Tibshirani (2011) uses $\mathbf{W}_d$ and selects a few variables using the Lasso penalty. However, this method fails when p is extremely larger than n.

Hence, the main objective of FC-SLDA is to find easily interpretable sparse discriminant direction with better performance in terms of speed and accuracy as compared with other competitive methods in the literature. This method is different from other methods that use $\mathbf{W}_d$, because it performs variable selection and classification simultaneously. The variables which are important for classification are retained. As a result, it provides more accurate results as compared with its competitive methods.

A general method of FC-SLDA was developed to find the column vectors of the discriminant transformation matrix $\mathbf{A}$ simultaneously. However, the general method can be computationally expensive, so we proposed an efficient sequential method to find each discriminant vector iteratively.

Different high-dimensional real data sets were used for illustrating performance of the methods, and they are compared with other competitive existing methods based on classification error and speed. The results show that FC-SLDA performs well when compared with other methods under fixed level of sparsity. It estimates the discriminant vectors sequentially, i.e., it uses a stepwise estimation method and it is faster than other methods that use $\mathbf{W}_d$. More interestingly,

the simplified version of our function constrained sparse LDA without the eigenvalue (FC-SLDA2) was the fastest method of discrimination though it performs with relatively higher classification error. Because this method selects very few variables but selects the important variables, the objectives of accuracy, sparsity and interpretability for high dimensional LDA are achieved.

In Chapter 5, we have proposed another interesting alternative method called sparse LDA using CPC (SDCPC) for high-dimensional classification problems. As we can see from Table 7.1, SDCPC assumes that the group covariance matrices have the same eigenvectors but different eigenvalues. These are weaker assumptions than those made by FC-SLDA and FC-SLDA2. This method performs effective classification for both $n > p$ and $p >> n$ data. SDCPC uses a modified stepwise estimation method and we imposed the cardinality constraint to find sparse discriminant vectors. It is an efficient estimation method for selecting common components iteratively. Moreover, it is computationally efficient, and it produces interpretable discriminant functions. As we can see in Table 7.2, SDCPC performs favorably compared to existing methods.

We know that, the traditional LDA works when $n > p$ and when all group covariance matrices are equal. However, in real world problem, group covariance matrices are, in general, not equal unless the groups come from the same population. SDCPC fills the gap that exits in classification problems involving unequal group-covariance matrices. A cardinality penalty is used to achieve sparsity. This penalty can help to select a few variables from a huge number of variables. From the numerical results using real data sets, sparse LDA based on CPC performs

well. Furthermore, our newly proposed method is compared with two other existing methods using real data sets. In general, SDCPC enjoys advantages in several aspects, including computational efficiency, interpretability, and an ability in identifying important variables for classification.

In Chapter 5, we also proposed another alternative discrimination method called sparse LDA using proportional cpc (SD-PCPC) for high-dimensional discrimination. This method assumes that group covariance matrices are proportional to each other. This method can be considered as an extension of SD-CPC and it is an ideal method when group covariances are proportional to each other. The proportional CPCs can be estimated using maximum liklihood or least squares method. We used the least squares method to estimate the CPCs in this particular method. We applied SD-PCPC on high-dimensional real data sets and we found that it performed better than other existing methods, especially when number of groups was not large.

In Chapter 6, we have proposed a new formulation of sparse LDA that is based on optimal scoring (OS). We refer to this method as SLDA-OS. We recall from Chapter 2 that binary discriminant analysis can be recast as regression analysis. Moreover, Clemmensen et al. (2011) proposed sparse discriminant analysis based on optimal scoring for classification problems with multiple groups. SLDA-OS assumes that all group covariance matrices are equal and it can be used for multi-group or binary classification problems. The method is similar to the Dantzig selector formulation for regression analysis. It is derived by considering the group indicators as dummy response variables. Because the Dantzig selec-

tor gives sparser results than the Lasso penalty and other sparsity penalties, it is an ideal method for a classification problem with an extremely large number of variables. That is, it selects a few useful variables from a huge number of variables. We applied SLDA-OS to both simulated and real data sets. We can see from the results in Table 7.2 that SLDA-OS performs better than the other methods in high-dimensional classification. In particular this method was found to be the most effective method for binary classification.

Results from the work with the six real data sets are presented in Table 7.2. We can see from the table that SDCPC, SLDA-OS, and SDA perform equally in classifying the Iris data with an MCE of 3%. They are followed by FC-SLDA and FC-SLDA2 with an MCE of 3.3% and 3.80%, respectively. The PLDA performed worst with an MCE of 4%. Therefore, we conclude that SDCPC, SLDA-OS, and SDA seem effective in classifying observations when the number of variables is less than the number of observations. At the same time, Fisher's LDA is a little better at classifying the Iris data, with an MCE of 2%. Similarly, when we compare the performances of the 7 methods in classifying the Rice data, SDCPC was found the best classifier with an MCE of 35.48%. Though an MCE of 35.48% is a poor classification performance, SDCPC performs better than the other 6 methods. The groups in the Rice data are very tight, which is why the 7 methods perform poorly in classifying the observations. Further, we can see from Table 7.2 that SD-PCPC was also found the best method in classifying the IBD data, with an MCE of 23.10%. It is followed by SDCPC with an MCE of 23.50%. The relatively better classification accuracy of SD-PCPC in classifying the IBD data set

is due to the fact that the group covariance matrices of IBD data are approximately proportional to each other. However, this method is the poorest method in classifying the Ramaswamy data, with an MCE of 48.15%. Therefore, we conclude that SD-PCPC performs better than other methods when group covariance matrices are proportional, but the number of groups should not be very large. SDCPC was found to be the best method in classifying the Leukemia data with an MCE of 13.17%. This method is effective in classifying observations when the group covariance matrices have the same eigenvectors but different eigenvalues. When we compare the performance of the 7 methods in classifying the Ovarian cancer data, SLDA-OS showed an extraordinary classification performance with just an MSE of 5.10% which is far better than the other methods. The Ovarian cancer data has only two groups and it may be that SLDA-OS is especially good at classifying a dataset that has just two groups. This should be examined in further work. Finally, when we see the performance of the 7 methods in classifying the Ramaswamy data, FC-SLDA was found to be the best method with an MCE of 13.13%. We know that the number of groups in Ramaswamy data is 14. Hence, we conclude that FC-SLDA seem to be the best method in classifying high-dimensional data with a large number of groups. FC-SLDA2 also performed well in classifying high-dimensional data sets and had the notable good quality of speed. This method was found to be the fastest method for classifying high-dimensional data sets. Therefore, FC-SLDA2 is recommended in classifying high-dimensional data sets if it is appropriate to compromise accuracy for speed.

Table 7.2: *Misclassification rate (in %) and time (in seconds) of seven sparse discriminant analysis methods on six real data sets.*

Data	FC-SLDA2		FC-SLDA		SDCPC		SD-PCPC		SLDA-OS		SDA		PLDA	
	Error	Time	Error	Time	Error	Time	Error	Time	Error	Time	Error	Time	Error	Time
Iris	3.80	0.0012	3.30	0.0013	3.0	0.0019	4.0	0.0013	3.0	0.0013	3.0	0.0013	4.00	0.0120
Rice	37.67	0.0050	37.00	0.0070	35.48	0.0068	37.21	0.0059	36.20	0.0068	37.15	0.0070	38.00	0.0760
IBD	34.63	97.5023	33.50	120.65	23.50	105.3508	23.10	155.3122	30.00	121.0200	30.65	112.2230	34.50	131.0600
Leukemia	31.42	18.2745	22.09	35.3201	13.17	53.1992	17.17	68.01289	21.50	19.6289	27.65	19.9700	27.33	35.2000
Ovarian	21.05	55.0350	19.03	59.1958	19.33	18.2347	18.21	21.0011	5.10	55.31280	19.31	58.3452	20.65	60.1024
Ramaswamy	18.00	109.3400	13.13	115.1903	32.50	118.4381	48.15	139.1301	16.33	113.1340	16.16	116.5012	–	–

7.2 Future research

Research is a continuous process where one idea brings forth another. Hence, every conclusion can be the beginning of new research. Therefore, our research could lead to further research on high-dimensional data. Many of the methods reviewed in Chapter 3 can be extended. For example, a ROAD to classification in high-dimensional space (Fan et al., 2012) can be extended to classification problems with multiple groups. Similarly other methods can further be improved. When we come to our contributions on sparse discrimination for high-dimensional problem, there are some nice ideas introduced in Chapters 4, 5, and 6 that can be further extended. For example, the fastest sparse LDA (FC-SLDA2) which was proposed in Chapters 4 can be extended by regularizing the within-groups matrix so as to find more accurate results. We know that most of the existing sparse discrimination methods are very slow, and they do not even work when p gets very large. Therefore, FC-SLDA2 is superior to the exiting methods in terms of speed. But it could be further extended to get more accurate results while it stays faster.

The SDCPC method which was proposed in Chapter 5 has the attractive features that it does not need equal group covariance matrices, although it does assume that group covariance matrices have common eigenvectors. Under this assumption, we have seen that SDCPC performs well in high-dimensional classification problems. If all of the group covariance matrices are proportional to each other, we have sparse discrimination with proportional CPC, called SD-

PCPC. This method might be further extended to the discrimination problem where some of the group covariance matrices are proportional while the remaining covariance matrices are not proportional.

We believe that our contributions of sparse LDA methods are possible alternatives for high-dimensional classification problems. They perform classification effectively and produce interpretable discriminant functions. But, they can also be used as a basis for further improvements and extensions of sparse discriminant analysis methods for high-dimensional data.

Bibliography

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

Bi, J., Bennett, K., Embrechts, M., Breneman, C., and Song, M. (2003). Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 3(Mar):1229–1243.

Bickel, P. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10:989–1010.

Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78.

Bouveyron, C., Girard, S., and Schmid, C. (2007). High-dimensional discriminant analysis. *Communications in Statistics,Theory and Methods*, 36(14):2607–2623.

Breiman, L. and Ihaka, R. (1984). *Nonlinear Discriminant Analysis Via Scaling and ACE*. Technical report 40. Department of Statistics, University of California.

Cai, D., He, X., and Han, J. (2008). Srda: An efficient algorithm for large-scale discriminant analysis. *Knowledge and Data Engineering*, 20(1):1–12.

Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 106:1566–1577.

Candès, E. J. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313–2351.

Clemmensen, L., Hastie, T., Witten, D., and Ersbøll, B. (2011). Sparse discriminant analysis. *Technometrics*, 53:406–413.

Clemmensen, L. K. H. (2013). On discriminant analysis techniques and correlation structures in high dimensions. Technical report, Technical University of Denmark.

Conrads, T. P., Zhou, M., III, E. F. P., Liotta, L., and Veenstra, T. D. (2003). Cancer diagnosis using proteomic patterns. *Expert Review of Molecular Diagnostics*, 3(4):411–420.

Dhillon, I. S., Modha, D. S., and Spangler, W. S. (2002). Class visualization of high-dimensional data with applications. *Computational Statistics and Data Analysis*, 41:59–90.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American statistical association*, 97(457):77–87.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

Fan, J. and Fan, Y. (2008). High dimensional classification using features annealed independence rules. *Annals of statistics*, 36(6):2605.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147:186–197.

Fan, J., Feng, Y., and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society, B*, 74:745–771.

Filzmoser, P., Gschwandtner, M., and Todorov, V. (2012). Review of sparse methods in regression and classification with application to chemometrics. *Journal of Chemometrics*, 26(3-4):42–51.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–184.

Flury, B. (1988). *Common Principal Components and Related Multivariate Models*. Wiley, New York.

Flury, L., Boukai, B., and Flury, B. D. (1997). The discrimination subspace model. *Journal of the American Statistical Association*, 92(438):758–766.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175.

Godbole, S. and Sarawagi, S. (2004). Discriminative methods for multi-labeled classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 22–30. Springer.

Guo, Y., Hastie, T., and Tibshirani, R. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86–100.

Haber, R., Rangarajan, A., and Peter, A. M. (2015). Discriminative interpolation for classification of functional data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 20–36. Springer.

Hage, C. and Kleinsteuber, M. (2014). Robust pca and subspace tracking from incomplete observations using ℓ_0-surrogates. *Computational Statistics*, 29(3-4):467–487.

Han, F., Zhao, T., and Liu, H. (2013). Coda: High dimensional copula discriminant analysis. *Journal of Machine Learning Research*, 14(Feb):629–671.

Hastie, T., Buja, A., and Tibshirani, R. (1995). Penalized discriminant analysis. *The Annals of Statistics*, 23:73–102.

Hastie, T., Tibshirani, R., and Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, 89(428):1255–1270.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417.

James, G. M., Radchenko, P., and Lv, J. (2009). Dasso: connections between the Dantzig selector and lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):127–142.

Johnson, R. A. and Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis*. Prentice-Hall, Upper Saddle River,NJ.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer-verlag, New York, 2nd edition.

Jolliffe, I. T., Trendafilov, N. T., and Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*, 12:531–547.

Krzanowski, W. J. (1999). Antedependence models in the analysis of multi-group high-dimensional data. *Journal of Applied Statistics*, 26:59–67.

Krzanowski, W. J., Jonathan, P., McCarthy, W. V., and Thomas, M. R. (1995). Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data. *Journal of the Royal Statistical Society. Series C*, 44:101–115.

Lachenbruch, P. (1975). *Discriminant Analysis*. The University of Michigan.

Mai, Q., Yang, Y., and Zou, H. (2015). Multiclass sparse discriminant analysis. *arXiv preprint arXiv:1504.05845*.

Mai, Q., Zou, H., and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99(1):29–42.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, London.

Marshall, A. and Olkin, I. (1979). *Inequalities: Theory of Majorization and Its Applications*. Academic Press, London.

MATLAB (2011). *MATLAB R2011a*. The MathWorks, Inc, New York.

McLachlan, G. (2004). *Discriminant Analysis and Statistical Pattern Recognition*, volume 544. Wiley. com.

Merchante, L. F. S., Grandvalet, Y., and Govaert, G. (2012). An efficient approach to sparse linear discriminant analysis. *arXiv preprint arXiv:1206.6472*.

Ng, M., Li-Zhi, L., and Zhang, L. (2011). On sparse linear discriminant analysis algorithm for high-dimensional data classification. *Numerical Linear Algebra with Applications*, 18:223–235.

Osborne, B. G., Mertens, B., Thomson, M., and Fearn, T. (1993). The authentication of basmati rice using near infrared spectroscopy. *Journal of Near Infrared Spectroscopy*, 1:77–83.

Pang, H. and Tong, T. (2012). Recent advances in discriminant analysis for high-dimensional data classification. *Journal of Biometrics & Biostatistics*.

Qiao, Z., Zhou, L., and Huang, J. Z. (2009). Sparse linear discriminant analysis with applications to high dimensional low sample size data. *International Journal of Applied Mathematics*, 39(1):48–60.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P., et al. (2001). Multiclass cancer

diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98(26):15149–15154.

Ramey, J. A. and Young, P. D. (2013). A comparison of regularization methods applied to the linear discriminant function with high-dimensional microarray data. *Journal of Statistical Computation and Simulation*, 83(3):581–596.

Rao, C. (1952). *Advanced Statistical Methods in Biometrics research*. John Wiley & Sons.

Rencher, A. (1992). Interpretation of canonical discriminant functions, canonical variates, and principal components. *The American Statistician*, 46:217–225.

Rencher, A. C. (2002). *Methods of multivariate analysis*. John Wiley & Sons.

Seber, G. A. F. (2004). *Multivariate Observations*. Wiley, New Jersey, 2nd edition.

Shao, J., Wang, Y., Deng, X., and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics*, 39(2):1241–1265.

Sharma, A. and Paliwal, K. K. (2008). A gradient linear discriminant analysis for small sample sized problem. *Neural Processing Letters*, 27(1):17–24.

Srivastava, M. S. and Kubokawa, T. (2007). Comparison of discrimination methods for high dimensional data. *J. Japan Statist. Soc*, 37(1):123–134.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of Royal Statistical Society*, 58:267–288.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to dna microarrays. *Statistical Science*, pages 104–117.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:91–108.

Trendafilov, N. T. (1994). A simple method for Procrustean rotation in factor analysis using majorization theory. *Multivariate Behavioral Research*, 29:385–408.

Trendafilov, N. T. (2010). Stepwise estimation of common principal components. *Computational Statistics and Data Analysis*, 54:3446–3457.

Trendafilov, N. T. (2013). From simple structure to sparse components: a review. *Computational Statistics*, Special Issue: Sparse Methods in Data Analysis, DOI:10.1007/s00180-013-0434-5.

Trendafilov, N. T. and Jolliffe, I. T. (2006). Projected gradient approach to the numerical solution of the SCoTLASS. *Computational Statistics and Data Analysis*, 50:242–253.

Trendafilov, N. T. and Jolliffe, I. T. (2007). DALASS: Variable selection in dis-

criminant analysis via the LASSO. *Computational Statistics and Data Analysis*, 51:3718–3736.

Trendafilov, N. T. and Vines, K. (2009). Simple and interpretable discrimination. *Computational Statistics and Data Analysis*, 53:979–989.

Vichi, M. and Saporta, G. (2009). Clustering and disjoint principal component analysis. *Computational Statistics and Data Analysis*, 53:3194–3208.

Wang, C., Cao, L., and Miao, B. (2013). Optimal feature selection for sparse linear discriminant analysis and its applications in gene expression data. *Computational Statistics and Data Analysis*, 66:140 – 149.

Wen, Z. and Yin, W. (2013). A feasible method for optimization with orthogonality constraints. *Mathematical Programming*, 142(1-2):397–434.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society, B*, 73:753–772.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation. *Biostatistics*, 10:515–534.

Wu, M. C., Zhang, L., Wang, Z., Christiani, D. C., and Lin, X. (2009). Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics*, 25(9):1145–1151.

Ye, J. (2005). Characterization of a family of algorithms for generalized discrimi-

nant analysis on undersampled problems. *Journal of Machine Learning Research*, pages 483–502.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Zou, M. (2006). Discriminant analysis with common principal components. *Biometrika*, 93:1018–1024.